Evaluating Faithfulness of LLM Generated Explanations for Claims: Are Current Metrics Effective?

Analysing the Capabilities of Evaluation Metrics to Represent the Difference Between Generated and Expert-written Explanations

Borislav Kolev Marinov[1]

Supervisor(s): Pradeep Murukannaiah, Shubhalaxmi Mukherjee

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Borislav Kolev Marinov
Final project course: CSE3000 Research Project
Thesis committee: Pradeep Murukannaiah, Shubhalaxmi Mukherjee, Xucong Zhang

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Large Language Models (LLMs) are increasingly used to generate fact-checking explanations, but evaluating how faithful these justifications are remains a major challenge. In this paper, we examine how well four popular automatic metrics—G-Eval, UniEval, FactCC, and QAGs—capture faithfulness compared to expert-written explanations. We look at how these metrics agree with each other, how they correlate with explanation similarity, and how they respond to controlled errors.Our findings show that while some metrics like UniEval and FactCC show some sensitivity to noise and partial alignment with expert reasoning, none of them reliably catch hallucinations or consistently reflect true faithfulness. Manual analysis also reveals that metric behavior varies depending on the type and structure of the claim. Overall, current metrics are only moderately effective and often biased toward the style of LLM-generated text. This study points to the need for more reliable, context-aware evaluation methods and offers practical insights for improving how we measure faithfulness in fact-checking tasks.

1 Introduction

Large Language Models (LLMs) are central to generative AI and are increasingly used in high-stakes domains such as healthcare, policy-making, and journalism (Liu et al., 2023). In these settings, it is not enough to verify the truth of a claim. LLMs must also produce explanations that faithfully reflect the underlying evidence. Unfaithful justifications can reduce trust and spread misinformation, even when the claim itself is correct (Russo et al., 2023).

A range of metrics has been proposed to evaluate faithfulness. However, assessing how well these metrics capture explanation quality is as important as generating the explanations themselves—particularly in fact-checking tasks, where claims are labeled *True*, *Half-True*, or *False* based on supporting evidence.

Prior work has shown that commonly used metrics, such as BLEU or ROUGE, correlate poorly with human evaluations in explanatory settings (Feher et al., 2025), (Russo et al., 2023). Newer frameworks like G-EVAL (Liu et al., 2023) use chain-of-thought prompting to improve alignment with human judgment, though they may still favor fluent, generic text over evidence-grounded content. Meanwhile, FactCheckBench (Wang et al., 2023) offers multi-dimensional judgments through the detection of factual errors in generated texts.

These studies rarely compare generated outputs to expert-written justifications and are more focused on the explanation itself rather than the performance of the evaluation metrics. This paper addresses the gap by asking: **How well do current evaluation metrics reflect the faithfulness of LLM-generated fact-checking explanations compared to journalist-written ones?** We evaluate four metrics—G-EVAL, UniEval, FactCC, and QAGs—on LLM-generated explanations for PolitiFact[1] claims, using journalist-written justifications as references. Our evaluation is structured around three research questions and analyses:

- RQ1: Score Correlation Analysis: To what extent do faithfulness metric scores align with each other and reflect the accuracy of the LLM's predicted labels?
- RQ2: Semantic Similarity Comparison: How do metric scores relate to the semantic/lexical similarity between generated explanations and human-written justifications for the same claims?
- RQ3: Sensitivity to Perturbations and Masking: How robust are the metrics to controlled perturbations in explanations, such as the addition of unsupported or unrelated sentences?

2 Faithfulness in Fact-Checking

This section will address the main concepts and misconceptions in the field of fact-checking related to faithfulness and illustrate the context of the research in greater detail.

Faithfulness vs. Factual Accuracy

Faithfulness refers to how well an explanation reflects the provided evidence, regardless of whether the claim or statements are factually correct. An explanation may be factually correct but unfaithful if it introduces unsupported content, or factually incorrect but still faithful if it reflects faulty evidence. This study focuses on evaluating faithfulness as captured by automatic metrics.

Existing Metrics Exploration and Limitations

The challenge of evaluating faithfulness in explanations has led to the development of various automatic metrics. Traditional n-gram overlap metrics such as BLEU, ROUGE, and METEOR have

[1] https://www.politifact.com/

been used in summarization and generation tasks (Liu et al., 2023). However, studies such as (Feher et al., 2025) show that these metrics often correlate poorly with human assessments when applied to explanatory texts in fact-checking scenarios. They fail in capturing the variety of thought processes and text structures.

Classifier-based approaches like SummaC (Laban et al., 2022) use Natural Language Inference (NLI) models to assess sentence-level consistency. While it improves upon lexical metrics, it is domain-specific, relying on high-quality NLI training data, and can perform worse on long explanations.

A recent development is G-EVAL, which uses LLMs to generate evaluation scores through chain-of-thought reasoning. This has shown strong alignment with human judgment (Liu et al., 2023). However, G-EVAL has mainly been studied in summarization where its sensitivity to model-like phrasing raises concerns. It tends to prefer smooth, fluent phrasing rather than evidence-based content. Its applicability to fact-checking and faithfulness evaluation remains unexplored.

3 Experimental Setup

This section outlines the setting of the experiments and the resources that will be used during them. Subsection 3.1 presents the datasets, while Subsection 3.2 discusses the metrics which are examined within this study.

3.1 Datasets

This study makes use of the QuanTemp[2] and HoVer[3] datasets as main focal points of the experiments. In this section, we will outline the taken preprocessing steps for each dataset. The final structure of the datasets can be found in Appendix B.

3.1.1 QuanTemp

The QuanTemp dataset consists of fact-checking entries from diverse sources, but we restrict our analysis to entries from PolitiFact to ensure consistent structure and access to ground-truth journalist-written explanations. To avoid priming the LLM, we removed the original explanation and label from the evidence field. This filtering reduced the dataset from 1500 to 350 entries, as many lacked a clearly separated explanation section. Each remaining claim is labeled True, Half True, or False and categorized by taxonomy to support analysis in different types of claims (Viswanathan et al., 2024).

[2] https://www.avishekanand.com/projects/quantemp_project/

[3] https://hover-nlp.github.io/

3.1.2 HoVer

To complement our analysis on the QuanTemp dataset, we additionally constructed a sample from the HoVer dataset, which focuses on multi-hop fact verification. Unlike single-hop verification, which evaluates a claim based on a single piece of evidence, multi-hop fact verification requires combining information from multiple sources or documents to validate a claim. This often involves reasoning across several interconnected facts that span different Wikipedia articles (Jiang et al., 2020). The HoVer dataset introduces this complexity by annotating claims with evidence drawn from 2-hop, 3-hop, or 4-hop reasoning chains. Another thing to notice is that the claims in HoVer are a binary classification of truthfulness, compared to the tertiary one in QuanTemp, meaning the only labels used are True and False.

We select 50 entries per hop level and leverage HoVer's structured annotations, which specify Wikipedia article titles and sentence indices for each supporting evidence chain. Using this information, we extract the relevant sentences from the corresponding Wikipedia articles to reconstruct the full evidence required for multi-hop verification. Each sample is then reformatted to match the structure used in QuanTemp, pairing the claim with its evidence and label.

3.2 Metrics and Implementation

We have explored four metrics that are often used in the evaluation of NLP related tasks. We have made slight modifications to them to accommodate them in our experimental setting and the task of evaluating the faithfulness of fact-checking explanations.

3.2.1 G-Eval

G-Eval is an LLM-based evaluation framework that uses language models to assess LLM-generated text across criteria such as factual accuracy, consistency, coherence, and fluency—primarily in summarization and dialogue tasks (Liu et al., 2023). It uses Chain-of-Thought (CoT) prompting and a form-filling approach to guide the evaluation process, with performance typically measured by correlation with human annotations.

In this work, we adapt G-Eval to assess faithfulness in fact-checking explanations. While the original setup uses GPT-4, we instead use Llama 3.1 (8B) running locally via the Ollama[4] framework to enable lightweight and cost-free evaluation. The evaluation prompt is based on a one-time CoT prompt, from which we extract generic steps for evaluating faithfulness. These steps are reused as guidance in the actual evaluation prompt. All the prompts can be found in Appendix A. Each explanation receives a score from 1 (unfaithful) to 5 (fully faithful). Due to Ollama and Llama not exposing token probabilities, we set the temperature to 0.4 to reduce determinism and encourage variation in the model's output. This mitigates the tendency observed in the original G-Eval study, where the model would repeatedly assign the same score across runs. Additionally, we aggregate two generated scores to produce the final value.

3.2.2 FactCC

FactCC is a weakly supervised, BERT-based method for evaluating factual consistency between summaries and source texts. It detects inconsistencies using techniques like sentence negation, pronoun swaps, and noise injection, classifying summaries as FAITHFUL or UNFAITHFUL based on aggregated sentence-level predictions with confidence scores between 0 and 1 for that label (Kryściński et al., 2019).

As the original implementation is outdated and not fully available, we use a modern adaptation [5] based on the same principles. It has a 512-token input limit, so we split the evidence into parts, preserving sentence boundaries, and evaluate each explanation sentence separately against each evidence chunk. A sentence is faithful if deemed so for at least 1 context block.

For each sentence, the model returns a binary label and confidence score. We compute a final explanation score by aggregating:

- **Faithful confidence (FA)**: The average confidence of sentences labeled as faithful.
- **Unfaithful disbelief (UD)**: Average of 1 - confidence for sentences labeled as unfaithful.
- **Precision**: The proportion of sentences labeled as faithful.

The final score (between 0 and 1) is computed as a combination of these components: *Final Faithfulness Score* = *Precision* · *FA* + (1-*Precision*) · *UD*. Explanations are labeled *Faithful* if the score exceeds 0.5.

This aggregation approach enables a more fine-grained judgment over multi-sentence explanations while maintaining alignment with the principles of the original FactCC model. It also allows us to use the final score as a measurement of faithfulness for the effectiveness evaluation of the metric.

3.2.3 QAGs

QAGs is a faithfulness evaluation approach that was originally introduced for summarization tasks. It combines two components: question generation (QG) and question answering (QA). The goal is to assess how well a generated text aligns factually with a reference document. The core idea is to extract factual statements from the generated explanation and convert them into questions using a QG model. These questions are then answered using a QA model, with the reference document or supporting evidence as input. If the generated content is faithful, the evidence should allow accurate answering of the questions implied by the explanation (Wang et al., 2020).

We adapt a QAGs-style pipeline for the fact-checking task. For each sentence in the explanation, we generate a question using a pretrained model[6] and use a QA model[7] to extract answers from the evidence. Due to input length limits, the evidence is split into overlapping sentence-based chunks (max 512 tokens), with sentence carryover to preserve context.

For each question, the most relevant chunk is used to extract an answer. We then compute the cosine similarity between the original sentence and the retrieved answer using sentence embeddings[8]. If the answer closely matches the sentence, it is considered grounded. The final QAGs score is the average of the best sentence-level similarities, ranging from 0.0 (unfaithful) to 1.0 (faithful), and supports fine-grained, comparative evaluation.

3.2.4 UniEval

UniEval is a unified framework for evaluating various aspects of generation using task-specific prompts to guide scoring by fine-tuned LLMs. It achieves a versatility of evaluation capabilities by transforming every task in a Boolean Question Answering problem (Zhong et al., 2022).

[4] https://ollama.com/
[5] https://huggingface.co/manueldeprada/FactCC
[6] https://huggingface.co/valhalla/t5-small-qg-hl
[7] https://huggingface.co/deepset/roberta-base-squad2
[8] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

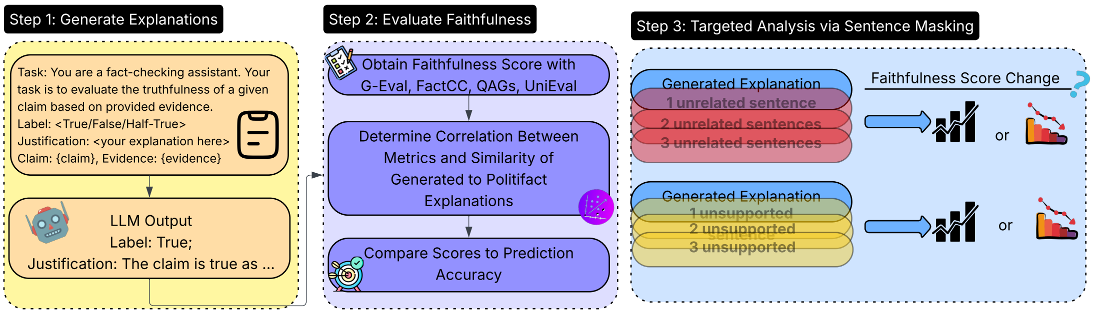


Figure 1: Experiment pipeline

We made minimal changes to the original evaluation pipeline and used the original pretrained model[9] that is optimized for factual consistency [10] . Like FactCC, this model has a token limit of 1024, requiring us to split the evidence into smaller chunks. Unlike FactCC, however, we evaluate the explanation as a whole against each chunk, and the final score is computed as the average across these evaluations.

3.3 Experimental Pipeline

Our experimental pipeline consists of three main stages, as depicted in Figure 1. In the first stage, we prompt a LLM (Llama 3.1 with 8 billion parameters) to generate both a label and a justification for a given claim based on associated evidence. This structured prompt mirrors a real-world fact-checking setting and outputs a natural language explanation of the model's reasoning.

In the second stage, we obtain scores from the four evaluation metrics presented in Subsection 3.2 to assess the faithfulness of these LLM-generated justifications. To objectively evaluate how well these metrics align with human-written justifications, we use a score that measures both lexical and semantic similarity between the generated and Politifact justifications. In the last phase, we conduct targeted tests to evaluate the ability of the metrics to capture faithfulness under manual hallucination addition to the explanation with both unrelated and unsupported sentences.

3.3.1 Comparison with Expert Explanations

To assess the degree to which metric scores reflect meaningful alignment with human-authored justifications, we compare each generated explanation with the corresponding journalist-written explanation. This comparison is based on two components:

- **Lexical F1 Score**: Measures overlap in lemmatized, stopword-filtered tokens.
- **Semantic Similarity**: We use sentence embeddings to compute the cosine similarity between the generated and reference explanation sentences. We compare each sentence in the generated explanation to all sentences in the reference explanation, and take the maximum similarity for each. The same embedding model used in the QAGS evaluation pipeline is employed here, as it is the most downloaded sentence embedding model on Hugging Face [11].

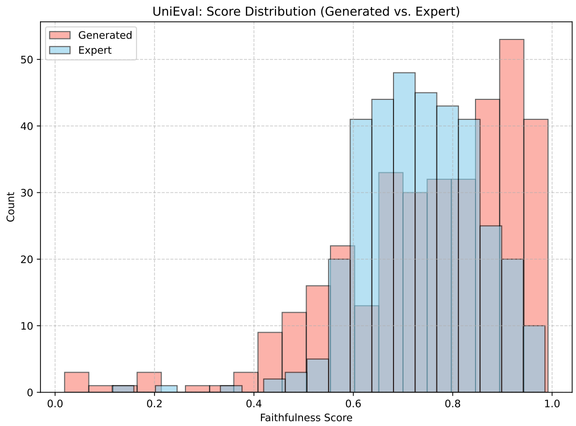
We compute a combined score using a weighted average of these two measures: *Combined Score* = $\alpha\cdot$ *Lexical F1* + $(1-\alpha)\cdot$ *Semantic Similarity*. This score provides a proxy for ground-truth similarity and is used to normalize and interpret metric scores. The weight (α) for lexical similarity is significantly smaller than semantic similarity, but is still included in order to capture potential use of the same pronouns and main pieces of evidence that could overlap. The weight $\alpha = 0.1$ was chosen based on manual inspection of explanation pairs. This value best represented perceived similarity, introducing a representative balance between the consistency of terminology usage and similarity in semantic meaning in reasoning. Our main hypothesis is that **higher similarity should lead to higher metric scores**.

3.3.2 Capturing Faithfulness

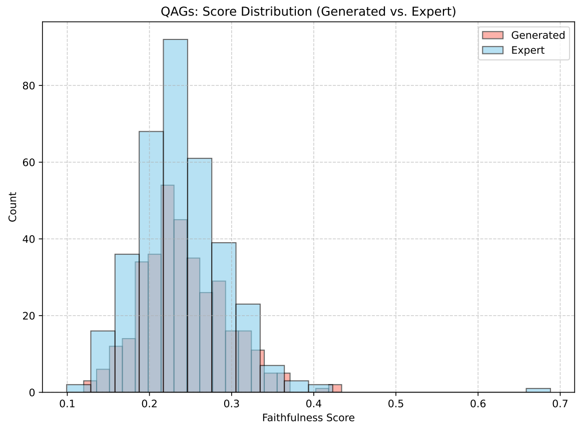
To evaluate whether existing metrics are capable of reliably capturing faithfulness, we conduct six targeted tests divided into two categories.

(1) Unrelated Noise Addition: We add 1, 2, and 3 sentences of unrelated information in steps to

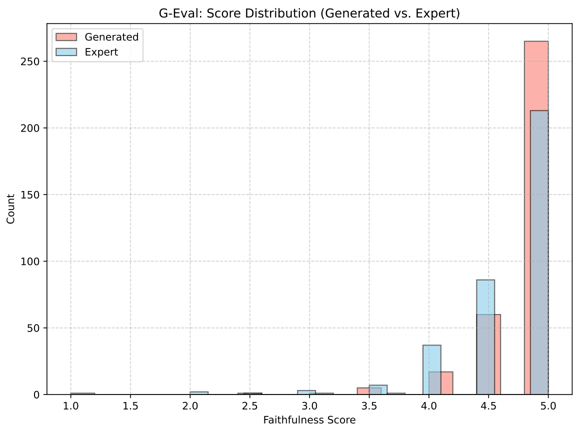
[9] https://github.com/maszhongming/UniEval
[10] https://huggingface.co/MingZhong/unieval-fact
[11] https://huggingface.co/



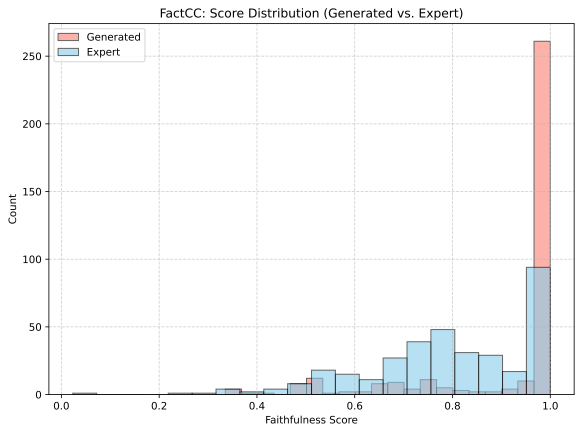
(a) UniEval



(b) QAGs



(c) G-Eval



(d) FactCC

Figure 2: Faithfulness score distributions for generated and expert-written explanations across four evaluation metrics. Each subplot (a–d) compares the metric's scoring behavior on LLM-generated outputs versus ground-truth justifications from journalists.

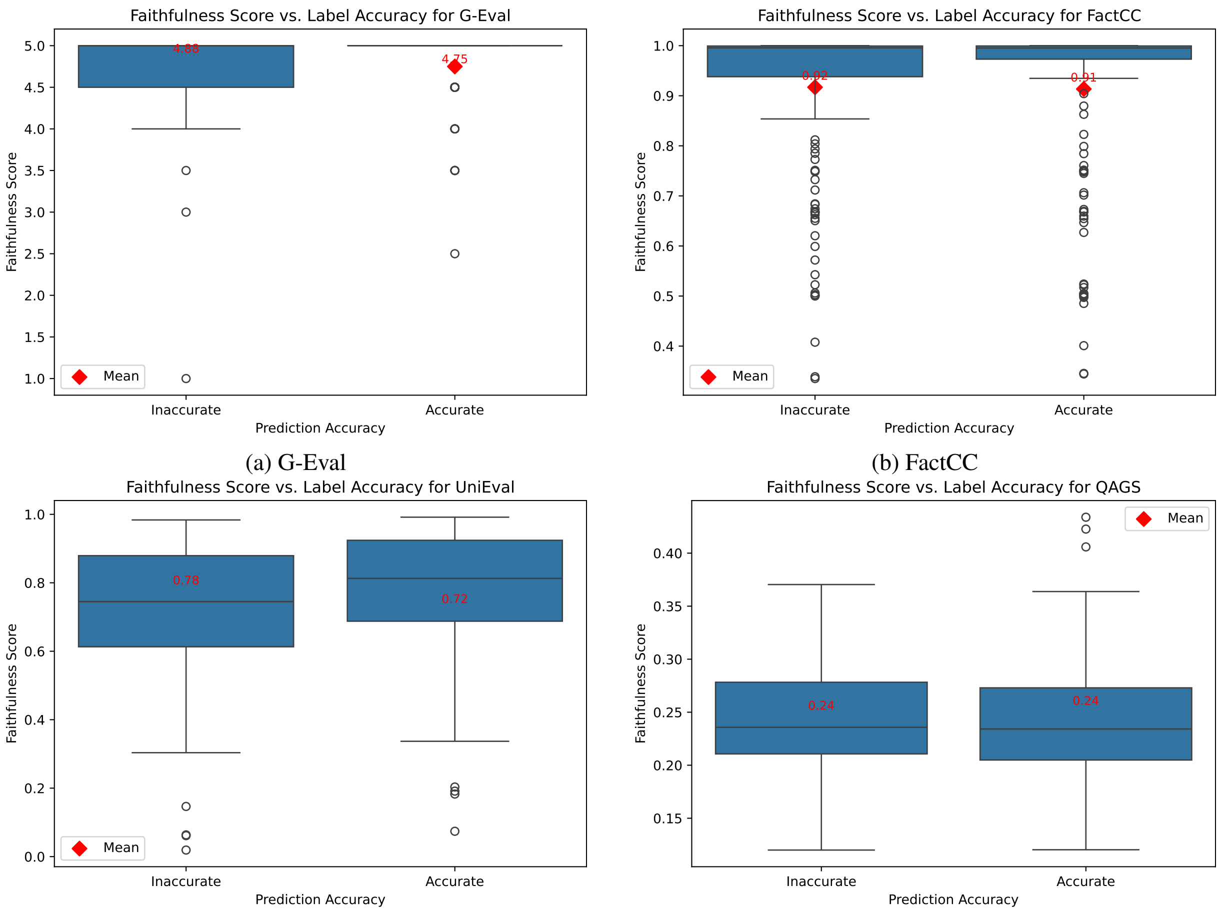
the end of the LLM-generated explanation. These sentences are semantically unrelated to the claim or evidence but syntactically coherent and factually accurate. This tests whether metrics can detect hallucinated off-topic content. Based on the known political context of the claims, we have a set of sentences to introduce randomly to the explanation. Examples of such sentences can be found in Appendix D.

(2) Related but Unsupported Information Addition: We perform the same steps as before but for sentences that are topically related to the claim but not supported by the given evidence. These sentences would appear relevant but cannot be grounded in the evidence, testing the metric's sensitivity to factually unsupported yet plausible hallucinations. The semantic similarity of each sentence on its own is computed against the evidence to ensure that it is not present. The sentences were generated with the help of LLMs for the set of generated explanations and the prompt used for them can be found in Appendix F.

Each of these six conditions is evaluated using the four metrics. Metrics that assign higher scores to explanations with minimal or no added unrelated/unsupported information are considered more reliable. This allows us to assess each metric's sensitivity to known variations and their robustness to superficial coherence.

4 Results

We first compare metric performance by analyzing score correlations and prediction accuracy per claim. Next, we assess how metric scores relate to differences between generated and PolitiFact justifications to test our hypothesis from Subsubsection 3.3.1. Finally, we evaluate the targeted tests to assess each metric's ability to reliably capture faithfulness.



(a) G-Eval (b) FactCC (c) UniEval (d) QAGs

Figure 3: Box plots showing the relationship between metric scores and label accuracy, with the mean score in red for each metric. This visualization helps assess how well each evaluation method differentiates between correctly and incorrectly predicted claims.

4.1 Metrics Comparison

The comparison between metrics consists of two stages. First, we examine how the different metrics relate to each other based on the scores they assign to both the generated and ground-truth explanations in the Quantemp dataset. The score distributions for ground-truth and LLM-generated explanations are shown in Figure 2. For ground-truth explanations, QAGs, UniEval, and FactCC display relatively well-distributed scores, in some cases resembling a Normal distribution. However, this pattern does not hold for UniEval and FactCC when applied to generated explanations, where scores are more heavily skewed towards the higher end. A consistent pattern is observed in QAGs' scoring behavior, which produces scores resembling a roughly normal distribution centered around 0.2, with a slight skew toward higher values for generated explanations. Compared to the other metrics, QAGs has more conservative scoring tendencies, assigning noticeably lower scores overall. In contrast, G-Eval tends to overestimate the faithfulness of the generated explanations, showing low variance, while it displays greater variation in scores for expert explanations, assigning them lower scores more frequently.

Secondly, we examine whether any metric demonstrates a meaningful correlation between high faithfulness scores and accurate label predictions by the LLM. As shown in Figure 3, the mean faithfulness scores (marked in red) are often slightly higher for inaccurately predicted labels, indicating weak or even counterintuitive alignment. However, the median scores are consistently higher for accurately predicted cases across all metrics, suggesting that higher faithfulness scores are more common among better label predictions. This highlights the nuanced and sometimes inconsistent behavior of the metrics and motivates further analysis of their reliability and sensitivity.

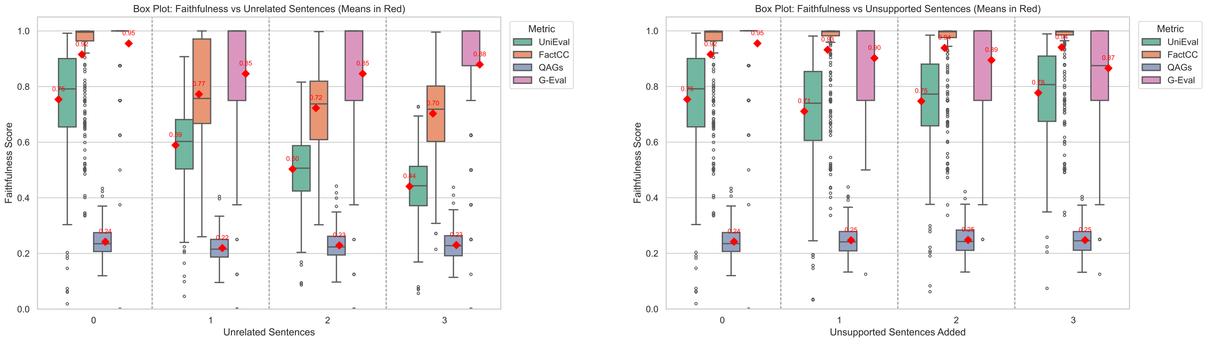


Figure 4: Faithfulness score distributions across evaluation metrics for (left) LLM-generated explanations with unrelated sentence perturbations and (right) unsupported factual sentences. Red diamonds indicate mean scores per metric per perturbation level.

	G-Eval	FactCC	UniEval	QAGs	Similarity
G-Eval	-	0.06 (0.29)	**0.09** (0.10)	-0.10 (0.07)	0.04 (0.47)
FactCC	0.06 (0.29)	-	-0.12 (0.02)	-0.04 (0.41)	**0.23** (<0.01)
UniEval	**0.09** (0.10)	-0.12 (0.02)	-	-0.05 (0.38)	-0.09 (0.09)
QAGs	-0.10 (0.07)	-0.04 (0.41)	-0.05 (0.38)	-	**0.15** (0.007)
Similarity	0.04 (0.47)	**0.23** (<0.01)	-0.09 (0.09)	**0.15** (0.007)	-

Table 1: Spearman (ρ) correlations between metric scores and similarity with p-values in parentheses. Bold indicates the highest value per row.

4.2 Ground Truth Alignment

In general, we observe that none of the metrics demonstrate a strong Spearman correlation with the computed similarity scores. In Table 1, no value exceeds 0.23. Among them, FactCC and QAGs show the highest correlations, offering a subtle yet tangible indication in support of the hypothesis from Subsubsection 3.3.2 that higher similarity between generated and reference explanations aligns with relatively higher faithfulness scores. We can also observe that these relations are statistically significant, as their p-values are the only ones that less than 0.05. By contrast, UniEval and G-Eval show low or even negative correlations with similarity, suggesting they are prone to giving lower scores to human-like writing and reasoning, even when generated by an LLM. This reinforces the results from the previous section about their scoring tendencies.

We use Spearman's rank correlation because it doesn't assume a linear relationship between scores and similarity. It's also more reliable when dealing with skewed or bounded data, like in our case. Since it measures how consistently one set of values increases with another, it's a good fit for our goal: checking whether metrics reliably give higher scores to explanations that are more semantically similar (De Winter et al., 2016).

4.3 Capturing Faithfulness Results

Figure 4 presents the distribution of faithfulness scores under controlled perturbations for each evaluation metric. Across unrelated perturbation, we observe that UniEval and FactCC display noticeable score degradation as noise increases, demonstrating some degree of sensitivity to faithfulness violations. QAGs, by contrast, maintains consistently low scores across all cases. G-Eval shows higher starting scores and a smaller drop, indicating that it may be more robust but potentially less sensitive to fine-grained factual inconsistency.

We can determine that the metrics are able to recognize the presence of the unrelated sentences, as the scores drop significantly between the original explanation and the changed one. On the other hand, unsupported but related sentences are a bigger challenge, as the overall scores do not show any clear indication that the explanation has deteriorated in faithfulness.

5 Discussion

Across our experiments, all the metrics show some alignment with the expectations we outlined in Section 3. However, this alignment is not consistent across metrics, nor is it strong enough to confidently say that current methods are effective at evaluating faithfulness. From the boxplots in Figure 4, we can see that FactCC and UniEval respond clearly to unrelated content, with noticeable and consistent score drops as hallucinations are introduced. G-Eval, on the other hand, tends to assign high scores even when faithfulness should decrease, suggesting it may be less sensitive to subtle inconsistencies. QAGs, in contrast, consistently gives lower scores, which might reflect a more cautious scoring approach that has difficulties in providing distinctive scores between explanations.

We also find that both UniEval and G-Eval tend to favor LLM-generated explanations, even when compared to expert-written ones (Figure 2). This matches findings from the original G-Eval paper, where the metric aligned more with LLM explanations than crowd worker annotations.

Beyond the quantitative findings, we conduct a targeted manual inspection of explanations where the similarity to expert justifications was high (similarity > 0.7) but the assigned metric score was low (< 0.5) within the QuanTemp dataset. These cases reveal specific limitations and behavioral tendencies of the metrics (Appendix E). For G-Eval, we observe that even when explanations closely mirror expert reasoning, the metric penalizes explanations that lead to incorrect conclusions. For instance, in statistical claims, a faithful but logically mistaken explanation receives a low score, highlighting G-Eval's emphasis on interpretation rather than solely faithfulness. In contrast, FactCC shows sensitivity to comparison and conflicting claims, rewarding explanations that incorporated concrete numerical or exact evidence, suggesting its alignment with lexical cues rather than broader semantic fidelity. UniEval displays a preference for more complete, elaborated explanations, particularly in statistical and comparison claims. Compressed or under-specified justifications, even if correct, are often scored lower, especially for false claims, where full-fact inclusion seems essential for a high faithfulness rating.

We conduct an inspection for the HoVer dataset as well. A critical observation is that very small changes in a claim have drastic effects on the performance of the LLM. The addition of one word can make the LLM predict the wrong label even if the explanations are the same and their faithfulness scores are high. This highlights the overall inconsistency and limitation of LLMs and faithfulness evaluation metrics in fact-checking (Subsection E.4).

6 Conclusions and Future Work

Our results show that current faithfulness metrics offer partial insight into explanation quality, but lack consistent reliability. We observe that metric scores show weak correlation with both label accuracy and each other, highlighting limited agreement on what constitutes a faithful explanation between metrics. This trend is also reflected in the similarity analysis. Similarity to the reference explanation does not consistently lead to higher metric scores for the generated one—particularly for G-Eval and UniEval. FactCC and QAGs exhibit slightly stronger alignment with reference texts, but their correlations remain weak. On the other hand, UniEval and FactCC respond to added hallucinations with clearer score degradation, showing some sensitivity to faithfulness violations. Conversely, G-Eval tends to assign uniformly high scores, while QAGs remains consistently low, reflecting distinct biases in scoring behavior. These findings highlight the limitations of current metrics, particularly LLM-based ones like G-Eval, which show ambiguity in scoring criteria despite explicit prompts. Overall the final answers to all the research questions posed in this paper serve to reach a single conclusion – the evaluation metrics exhibit inconsistent and unreliable behavior in their faithfulness scores with their performance varying in quality across analyses. However, the positive aspects of the results show the need for further development and exploration alongside the advancements of LLMs.

Several possibilities remain open for exploration. Future work should expand the PolitiFact dataset for broader benchmarking against expert justifications. It is worth exploring and training the metrics and models for UniEval, FactCC and QAGs specifically for fact-checking. A deeper understanding of the sensitivity of the metrics can be explored through more varying perturbation tests that consider adding sentences in the middle of the explanations. Furthermore, applying stronger LLMs (e.g.GPT-4) to G-Eval can improve score variance and sensitivity. Finally, a critical step to expanding this research could be the participation of more human annotators that can help in the analysis of the similarity between explanations and faithfulness alignment.

7 Responsible Research and Reproducibility

This research aligns with the ACL Responsible NLP Research Checklist [12] by addressing ethical considerations, transparency in the use of AI, and ensuring the reproducibility of all experimental procedures and artifacts.

[12]https://aclrollingreview.org/static/responsibleNLPresearch.pdf

7.1 Limitations of the Work

Our study investigates the effectiveness of current automatic metrics in evaluating the faithfulness of fact-checking explanations generated by LLMs. While our analyses reveal where metrics succeed and fail, they remain approximations of human judgment. Faithfulness is context-dependent and inherently subjective in formulation. Thus, even when a metric correlates well with human-written justifications, this should not be interpreted as conclusive validation of its reliability in real-world applications. Furthermore, it is crucial to note that the models used in the experiments were not trained for the task of fact-checking and in the exact context of the datasets, which could allow for improvement of the performance of the metrics and explanation quality. The results of the research could have been influenced by the scale of the LLMs and models used, as they are commercially available models that can run on a local machine, so they are smaller than more advanced models such as GPT-4. Another limitation to be considered in relation to this is the fact that the experiments were performed on a Lenovo Legion 5 NVIDIA GeForce RTX 3060 laptop GPU, which affected the time needed for the experiments and potentially the final results.

7.2 Potential Risks and Misuse

Although our work does not involve sensitive personal data or interaction with users, it contributes to an area with societal implications: misinformation and automated fact-checking. There is a risk that high metric scores could falsely signal that a generated explanation is faithful, when it may still mislead readers. We caution against deploying these metrics without human oversight, particularly in domains such as news, healthcare, or political content.

7.3 Data Usage and Acknowledgement

We used a subset of the QuanTemp dataset (350 entries) and constructed a balanced subset of the HoVer dataset by selecting 50 claims for each reasoning depth: 2-hop, 3-hop, and 4-hop. All of the data from these datasets is intended to be used for research purposes, which is how it has been employed in our study.

7.4 Scientific Artifacts

We used publicly available datasets (QuanTemp, and HoVer) with citations, and created multiple scientific artifacts, including:

- Generated explanations using LLMs for all evaluated claims.
- Implementation of each metric (G-Eval, UniEval, FactCC, QAGs)
- Metric scores for each explanation
- Controlled perturbation datasets with sentence additions.
- Prompt templates and scoring scripts.

These artifacts are publicly available in a GitHub repository [13], alongside the code for their generation. These artifacts do not serve as a way to portray that a statement made in them is actually truthful or not. They are to be used solely in analysis of the performance of metrics in fact-checking scenarios and the scores obtained do not pertain to any validity of the information in the explanations. The time to create these articles was around 100 hours of GPU usage with G-Eval evaluations taking the longest of around 25 hours per run on the whole dataset.

7.5 Use of AI and LLMs

All explanations evaluated in this study were generated by large language models (e.g., LLaMA 3 8B via LangChain). G-EVAL scores were also derived from LLM outputs prompted to judge explanation faithfulness. We release the prompts, model versions, and decoding settings used to ensure transparency. In terms of the use of AI during the research itself, the prompts were coordinated with ChatGPT to be adjusted so they can serve as clear instruction for an LLM. Furthermore, AI was utilized in the final formulation of the code and paper, helping in writing comments and documentation to facilitate use of the code base. In terms of its use in the writing of the paper, it was used to generate possible pointers to improve different sections of the text, without actually writing anything and solely making suggestions that were considered during editing in some cases and disregarded in others. The structure of the prompts used can be found in Appendix G.

References

Joost C. F. De Winter, Samuel D. Gosling, and Jeff Potter. 2016. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. 21(3):273–290.

[13] https://github.com/bobecrec/llm-fact-checking-faithfulness-metrics

Darius Feher, Abdullah Khered, Hao Zhang, Riza Batista-Navarro, and Viktor Schlegel. 2025. Learning to generate and evaluate fact-checking explanations with transformers. *Engineering Applications of Artificial Intelligence*, 139:109492.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. *Preprint*, arxiv:2011.03088 [cs].

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *Preprint*, arxiv:1910.12840 [cs].

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Daniel Russo, Serra Sinem Tekiroglu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Venktesh Viswanathan, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. *Preprint*, arxiv:2403.17169 [cs].

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *Preprint*, arxiv:2004.04228 [cs].

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2023. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *Preprint*, arxiv:2210.07197 [cs].

A G-Eval Prompts

A.1 Generate Steps Prompt

You are a faithfulness evaluator.

Task: You are given a claim, a text that presents relevant evidence and an explanation that has been generated to explain if the claim is true or false. You are to rate the faithfulness of the provided explanation on a scale from 1 to 5.

Faithfulness Criteria:

*- Faithfulness means the explanation should be based *only* on the provided information in the evidence entry above.*

*- It must **not** hallucinate facts not present in the evidence. The facts used must be present in the provided evidence.*

- An explanation that is factually correct but irrelevant to the evidence is NOT faithful.

*- Give a **score from 1 to 5**, where 1 = not faithful at all, and 5 = fully faithful. Decimals are allowed.*

Generate 5 evaluation steps to perform this task successfully.

A.2 Evaluate Faithfulness Prompt

You are a faithfulness evaluator.
Task:
*Evaluate the *faithfulness* of the following explanation based on the provided claim and evidence.*
*- **Claim:** [claim]*
*- **Evidence:** [evidence]*
*- **Explanation:** [explanation]*
Evaluation Criteria:
*- Faithfulness means the explanation should be based *only* on the provided information in the evidence entry above.*
*- It must **not** hallucinate facts not present in the evidence. The facts used must be present in the provided evidence.*
- An explanation that is factually correct but irrelevant to the evidence is NOT faithful.
*- Give a **score from 1 to 5**, where 1 = not faithful at all, and 5 = fully faithful. Decimals are allowed.*
Use the following steps to reason about the score:
<steps>
Final Answer Format:
*At the end of your response, return your final score of the score **in this exact format and wording** (do not deviate):*
Final Faithfulness Score: [a number from 1 to 5]
Example:

Final Faithfulness Score: 3
Now begin your evaluation.

B Final Dataset Structure

These are the final structures for the QuanTemp and HoVer dataset.

B.1 QuanTemp Structure

- ***claim***: *"More than 50 percent of immigrants from (El Salvador, Guatemala and Honduras) use at least one major welfare program once they get here."*
- ***evidence***: *"The crisis at the border brought on by thousands of young people seeking entry, some with a parent and many without, has fueled an immigration debate that was already overheated ..."*
- ***PolitiFact explanation***: *"O'Reilly said that more than 50 percent of immigrants from El Salvador, Guatemala and Honduras use at least one welfare program. ...",*
- ***label***: *"True",*
- ***taxonomy label***: *"statistical"*

B.2 HoVer Structure

- ***claim***: *"ḦMS S̈ussexẅas an 80-gun third-rate ship, of the line of the English Royal Navy, for the Kingdom of England on the island of Great Britain from the 10th century2̆014when it emerged from various Anglo-Saxon kingdoms2̆014until 1707, when it united with Scotland to form the Kingdom of Great Britain.",*
- ***evidence***: *"On board were possibly 10 tons (330,000 troy oz) of gold coins. The Kingdom of England was a sovereign state on the island of Great Britain from the late 9th century, when it was unified from various Anglo-Saxon kingdoms, until 1 May 1707, when it united with Scotland to form the Kingdom of Great Britain, which would later become the United Kingdom.", "justification": "The evidence does not mention HMS S̈ussexät all. It only talks about the Kingdom of England and its unification with Scotland in 1707, but there is no information about a ship called HMS S̈ussex.̈",*
- ***label***: *"True",*
- ***num hops***: *"2"*

C Similarity vs Faithfulness Score per Metric

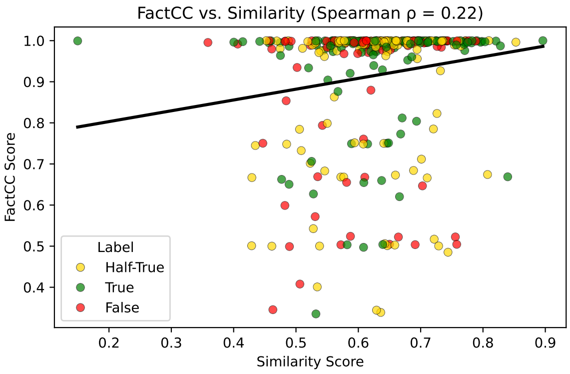


Figure 5: FactCC Scores vs. Similarity for Different Labels

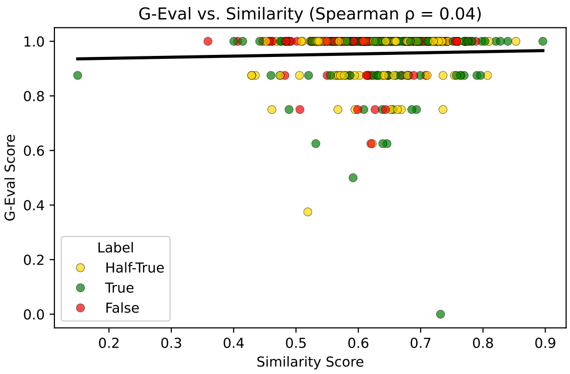


Figure 6: G-Eval Scores vs. Similarity for Different Labels

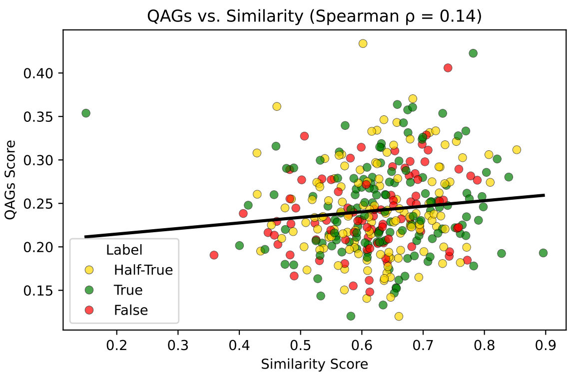


Figure 7: QAGs Scores vs. Similarity for Different Labels

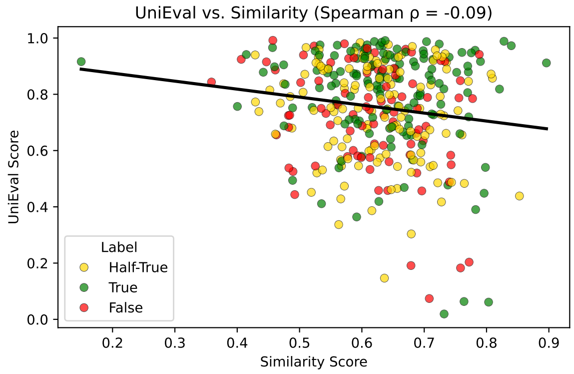


Figure 8: UniEval Scores vs. Similarity for Different Labels

D Unrelated Sentences Example

> *The moon has no atmosphere, which means there is no sound.*
>
> *Most dust in your home is made of dead skin cells.*
>
> *Octopuses have three hearts and blue blood.*
>
> *The Eiffel Tower can be 15 cm taller during the summer.*

E Manual Inspection Examples

E.1 G-Eval

Claim: A poll conducted in Egypt revealed that 80 percent of Egyptians oppose receiving foreign aid from the United States.

Expert Explanation: Barry Hinckley, in raising questions about whether the United States should be giving foreign aid to countries such as Egypt, said that 80 percent of Egyptians oppose receiving foreign aid from the United States. A Gallup poll taken in February shows that he's right in the ballpark. We rate his statement as True.

Generated Explanation: According to two independent polls conducted in Egypt, one by Gallup between Jan. 31 and Feb. 7 with a margin of error of plus or minus 3.4 percentage points, and another by the Pew Research Center from March 19 to April 10 with a margin of error of plus or minus 4.2 percentage points, more than 80 percent of Egyptians do not oppose receiving foreign aid from the United States.

Similarity Score: 0.73 (very similar)
G-Eval Score (Generated): 1
G-Eval Score (Expert): 4.5

E.2 FactCC

Claim: "Muslim nations did not 'call out' the people who celebrated the 9/11 attacks."

Expert Explanation: O'Reilly said that Muslim nations did not "call out" the people who celebrated the 9/11 attacks. So far as we can tell, there was no official condemnation of people celebrating the 9/11 attacks. However, Muslim governments, and religious leaders, condemned the attacks themselves, as did many average Muslims. O'Reilly's statement is accurate as far as it goes, but it leaves out important information that might lead someone to reach a different conclusion. That meets our definition of Half True.

Generated Explanation: The evidence suggests that there were no official condemnations of the people who celebrated the 9/11 attacks by Muslim nations, but it also shows that many governments and top religious leaders condemned the attacks themselves, which is a different statement. Additionally, average Muslims in some countries attended public events to voice sympathy for the victims and repudiate the attacks, indicating that not all Muslims supported the celebrations.

Similarity Score: 0.74
FactCC Score (Expert): 0.99
FactCC Score (Generated): 0.49

E.3 UniEval

Claim: Newark funds "80 percent of the school budget from local property taxes."

Expert Explanation: In a legislative hearing, Rice claimed that Newark funds "80 percent of the school budget from local property taxes." Of a $970 million budget, more than $106 million, or about 11 percent, is raised from local taxes. State aid, grants and entitlements account for most of the budget, at $775 million. That's nearly 80 percent. We rate the statement False.

Generated Explanation: According to budget documents, local taxes account for 11 percent of the Newark Public Schools' overall budget in the 2011–2012 fiscal year, which contradicts Sen. Rice's claim that they pay 80 percent of the school budget from local property taxes.

Similarity Score: 0.77
UniEval Score (Expert): 0.56
UniEval Score (Generated): 0.20

E.4 HoVer

Claim 1: `The MV Bessel ran aground at the second largest island in the Mediterranean in 1972.`

Justification: `The evidence states that the MV Bessel (previously Nicolo Martini) ran aground at Portoscuso, Sardinia in 1972, which is confirmed to be the second largest island in the Mediterranean Sea.`

Score: 0.92857
Predicted Original Label "True": true

Claim 2: `The MV Bessel ran aground at the second largest island in the Mediterranean in` **summer** `of 1972.`

Justification: `The evidence confirms that the MV Bessel (previously Nicolo Martini) did indeed run aground in 1972, and it occurred on the island of Sardinia, which is identified as the second largest island in the Mediterranean.`

Score: 0.9844
Predicted Original Label "False": false

F Unsupported Sentences Generation Prompt per Claim and Explanation

You are a unique fact generator.

Extract an entity or concept from this claim: [claim]

Generate three general factual sentences about it. These sentences should be:

- Truthful and relevant to the entity,

- Not found in the claim or this piece of evidence: [evidence],

- Varied in nature (e.g., background info, unrelated historical context, trivia, associated fields, culture, politics, etc.)

Avoid repeating the original claim or using information from the evidence. Keep them factual, broad, and not obviously connected to typical uses in news or argumentation.

Return in this format:

- Sentence 1:

- Sentence 2:

- Sentence 3:

G LLM Prompts Used During Research

"Provide a clear and concise description to this method that does [include method function here]: [include method signature for arguments and return value, if any]"

"Help with formatting this figure/table/section of my latex paper that portrays [include description]: [latex code]"

"Suggest adjustments to the prompt such that an LLM would be able to perform the task described in it successfully and follow any formatting requests in its response: [include prompt]"

"Given this paragraph [include paragraph], give some points for improvement in phrasing/structure considering [aim of the paragraph and main function in the paper]"