Evaluating Evidence Attribution in Generated Fact Checking Explanations

Anonymous ACL submission

Abstract

Automated fact-checking systems often struggle with trustworthiness, as their generated explanations can include hallucinations. In this work, we explore evidence attribution for fact-checking explanation generation. We introduce a novel evaluation protocol, citation masking and recovery, to assess attribution quality in generated explanations. We implement our protocol using both human annotators and automatic annotators, and find that LLM annotation correlates with human annotation, suggesting that attribution assessment can be automated. Finally, our experiments reveal that: (1) the best-performing LLMs still generate explanations with inaccurate attributions; and (2) human-curated evidence is essential for generating better explanations. Code and data are available here: ANONYMISED.

1 Introduction

The dissemination of online misinformation poses a significant threat to society, with consequences ranging from seeding skepticism and discrediting science, to endangering public health and safety. To verify online claims at scale, automated fact-checking systems have been proposed to classify claims based on their truthfulness (Guo et al., 2022). However, debunking by simply calling a claim "false" can lead to a backfire effect, whereby the belief in a false claim is further reinforced rather than hindered (Lewandowsky et al., 2012).

As such, there's growing research on generating textual explanations to justify the outcomes of fact-checking systems using large language models (LLMs) in various contexts (Eldifrawi et al., 2024; Wiegreffe et al., 2022). For automated fact-checking, the typical input is a claim and a list of retrieved evidence passages, from which a subset of evidence passages is selected and fed to the LLM for explanation generation. Table 1 presents a typical example of the fact-checking task with a generated explanation, and Figure 1 illustrates the process of generating these explanations.

Claim: In 2018, for every "share" that a photograph of a sick boy gets, Facebook will donate one dollar. ... **Evidence passage 9:** Facebook does not make donations based on the number of shares or comments a particular post or photo may receive. ... **Evidence passage 10 :** The post is an example of a long-running scam that is employed to generate user interaction by falsely claiming that shares or comments will help pay for a sick child's care. ... **Evidence passage 11 :** In 2015, the Better Business Bureau released a statement about such scams...
Veracity: False
Selected evidence: {9, 10, 11}
Generated Explanation: The claim that Facebook will donate a dollar for every "share" that a photograph of a sick boy receives is false. It is crucial to note ... a particular post or photo may receive [9]. Such posts are typically part of a long-running scam ... will financially contribute to the care of a sick child [10]. Though it might initially appear benign,..., which in turn empowers them to disseminate further deceptions [11].

Table 1: An example based on PolitiHop with claim c, veracity v, evidence passages E. X is the generated explanation that cites E.

A challenge in this task is that the explanations must provide accurate attribution of their key statements (Nakov et al., 2021). In journalistic fact-checking practice, journalists curate their sources and ensure the accuracy of every cited statement (Baker and Fairbank, 2022). Machine-generated explanations should this follow convention to be trustworthy. But even if generated explanations contain attribution, how do we assess whether the attribution is faithful to the source evidence? Current evaluation methods are primarily based on the Attribution to Identified Source (AIS) framework (Rashkin et al., 2023), whereby an important feature is "explicature": the meaning of

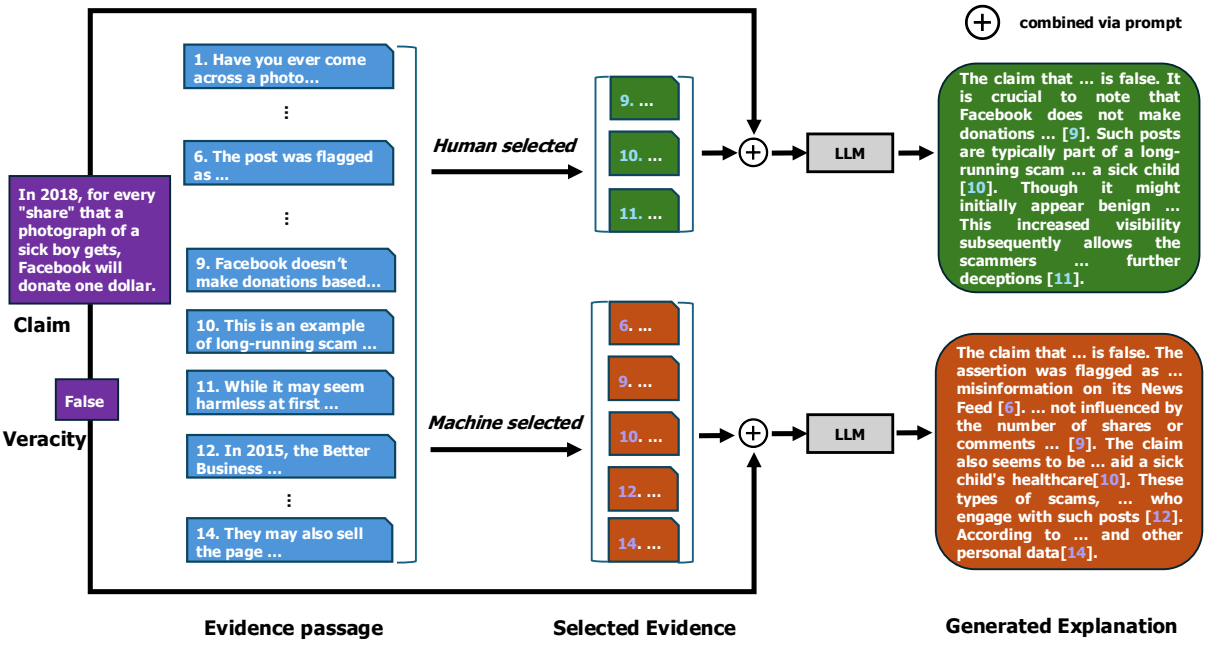

Figure 1: Automated explanation generation in fact-checking. Given a claim, its veracity and a list of evidence passages, a subset of these passages is selected, either by humans or machines, and input into a large language model (LLM) along with the claim to generate the explanation.

statement s in linguistic context c. Generating explicatures typically involves either costly human annotation or a non-trivial process of decontextualization, meaning it is rarely used. Consequently, most automated evaluation methods directly use Natural Lanugage Inference (NLI) between c and s to quantify AIS (Gao et al., 2023; Liu et al., 2023; Li et al., 2024).

In this work, we ground the notion of explicature by incorporating contextual information into the evaluation process. We introduce a novel evaluation protocol – *Citation Masking* and *Citation recovery* – to assess evidence attribution in generated explanations. The key idea is to mask the evidence citation in the generated explanation and ask annotators to recover them. The recovery process requires annotators to incorporate the linguistic context of a candidate statement during evaluation. This results in zero or more candidate statements, reflecting the complex nature of real-world attribution, and we hence frame it as a multi-label task. We first ask annotators to evaluate the attribution of a statement within the context of the full explanation. We then automate our protocol using NLI methods and LLMs as annotators. Additionally, we explore the influence of evidence selection on attribution quality, comparing human-curated and machine-selected evidence.

To summarise, our contributions are: (1) we introduce a new evaluation protocol for evidence attribution in generated explanations via citation masking and recovery; (2) we automate our protocol via LLMs and compare it with NLI models, showing that LLMs align better with human annotators; and (3) in terms of evidence source, we explore the influence of human- and machine-selected evidence on explanation generation and find that human-selected evidence produces explanations that are more accurate in terms of evidence attribution.

2 Related Work

Automated fact-checking aims to classify the veracity of a claim (Guo et al., 2022; Russo et al., 2023). However, debunking solely by providing a classification label (e.g. *False*) is neither effective nor convincing, and can induce a "backfire" effect where the erroneous belief is reinforced (Lewandowsky et al., 2012; Guo et al., 2022). This motivates the task of justification or explanation generation for fact-checking. Explanation generation in the literature is framed as either extractive or abstractive summarization over evidence (Atanasova et al., 2020; Kotonya and Toni, 2020; Xing et al., 2022). However, extractive methods struggle to produce explanations with sufficient context, and abstractive methods are prone to hallucination (Russo et al.,

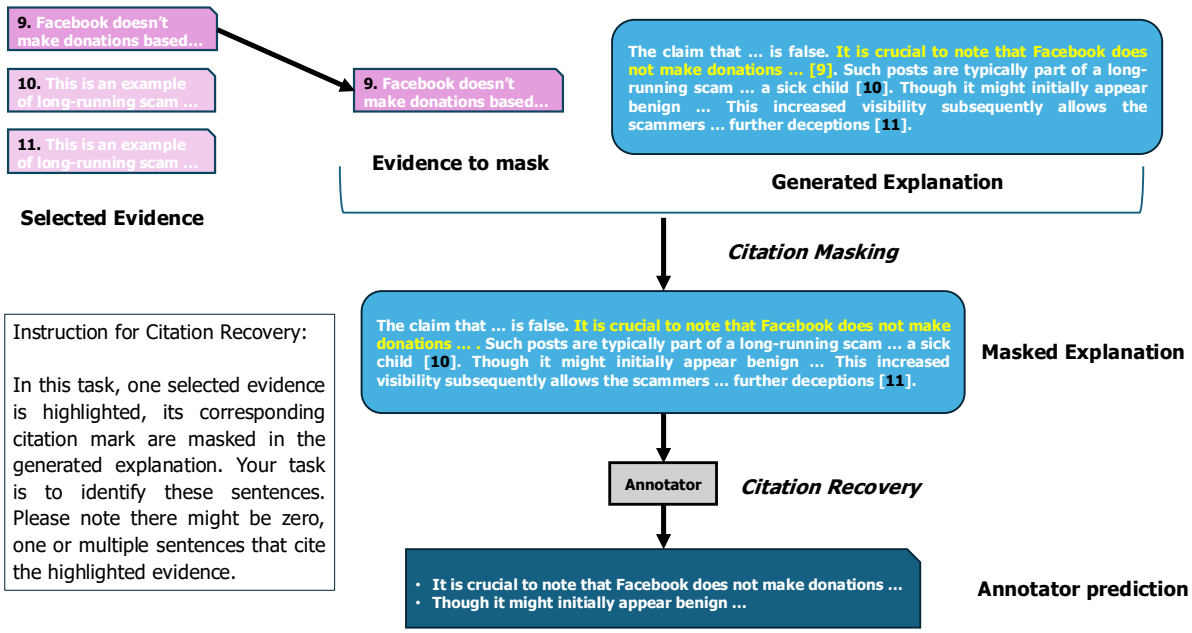

Figure 2: Our human assessment protocol: citation masking and citation recovery. Given the generated explanation and a list of evidence passages, we randomly masked one sentence e_k and mask its inline citation marker. Annotators are then required to perform citation recovery and predict the masked citation sentence.

2023).

To alleviate hallucination, most existing work incorporates citations in the generation and evaluates their attribution relative to sources (Liu et al., 2023; Gao et al., 2023; Yue et al., 2023; Huang and Chang, 2024; Li et al., 2024). Liu et al. (2023) introduced evaluation of citation quality, and used human judges to audit the verifiability of popular LLMs. Gao et al. (2023) developed automatic metrics along three dimensions – fluency, correctness, and citation quality – to assess LLMs for question answering. Yue et al. (2023) explored prompting LLMs and fine-tuning smaller LMs for automatic attribution evaluation. Most existing work is based on the Attribution to Identified Source (AIS) framework (Rashkin et al., 2023), where an annotator is asked whether the statement s in the explanation is fully supported by cited evidence c (*According to evidence c, s.*). An important assumption in AIS is explicature, an expanded statement that is interpretable out of context. Generating explicatures is a non-trivial process, often requiring costly human annotation to decontextualize statements so they can be interpreted in isolation. Most work directly adopts NLI without considering explicature (Liu et al., 2023; Gao et al., 2023). In this work, we attempt to bridge the gap by introducing a new evaluation protocol to evaluate evidence attribution in fact checking explanations.

3 Human Evaluation of Evidence Attribution

3.1 Dataset

We use PolitiHop (Ostrowski et al., 2021) as our dataset for all experiments. Table 1 shows an example instance, comprised of a **claim**, a **veracity** label, a list of **evidence passages** that provide additional relevant information about the claim, and **selected evidence** – a subset of passages selected by *human annotators* that is most relevant to the claim and collectively clarifies the veracity of the claim.[1]

3.2 Automatic Explanation Generation

PolitiHop does not provide an explanation for each claim, and as such we need to generate the explanations for our experiments. We developed an LLM-based explanation generation system which assumes the following as input: (1) a claim; (2) a list of evidence passages; and (3) the veracity of the claim (e.g. *False*). The rationale for assuming the veracity label as input is we see our explanation generation system being applied to a fact-checking

[1] In PolitiHop, the evidence in each claim can sometimes have multiple subsets (e.g. {8,10} and {6, 7, 8}); for these cases we randomly select one subset for our experiments.

Evi Src	Gen	Claim Len	Evi Size	Exp Len
Human	GPT-4	22.32	3.16	140.58
	GPT-3.5			
	LLaMA2			
Machine	GPT-4	22.32	5.07	171.26
	GPT-3.5		4.68	175.89
	LLaMA2		6.24	214.02

Table 2: Statistics of the data and generated explanations. The "Evi Src" column indicates whether evidence is human- or machine-selected. "Gen" means the generator model. "Claim Len" and "Exp (Explanation) Len" refer to token length, which is tokenized by OpenAI's tiktoken v0.7.0. Evi (Evidence) Size refers to the number of selected evidence passages.

classification system to provide a layer of interpretability in a real-world application. Because the set of evidence passages is typically large and not all of them are relevant, we propose a pipeline approach where we first do **evidence selection** and then feed the selected evidence, claim, and veracity label to an LLM to generate the explanation. Figure 1 illustrates the overall process.

Evidence Selection We experiment with two methods for evidence selection: (1) *Human*: we use the original human-selected evidence in Politihop; and (2) *Machine*: we (one-shot) prompt an LLM to select the most relevant evidence passages (see C14 for more details). Note that the evidence passages are numbered, which is important as the explanation cites them for key statements. We explore these two methods to understand the impact of this selection process – whether hand-curated evidence is necessary for explanation generation, or if machine-selected evidence is sufficient.

Explanation Generation Given a claim, veracity label, and selected evidence (machine- or human-selected), we (zero-shot) prompt an LLM to generate an explanation to clarify the claim. The prompt explicitly asks it to use in-line citations, making it clear that *all* given evidence passages should be used (see "Generated Explanation" in Figure 1; the full prompt is given in Appendix Table 12). We test a range of LLMs in our experiments, noting that we always use the *same* LLM for evidence selection and explanation generation. We generate explanations for 100 claims sampled from PolitiHop.

Models For the evidence selector and explanation generator, we experiment with three language models: GPT-4 (OpenAI, 2024), GPT-3.5, and LLaMA2 (Touvron et al., 2023).[2] Note that we always use the same LLM for both evidence selection and explanation generation, and as such the machine-selected evidence can also be interpreted as self-selected evidence. Table 2 presents data statistics for the explanation generation process. In general, we can see that the machine-selected evidence set tends to be larger (i.e. it includes more passages), and the explanations generated using machine-selected evidence are also longer.

3.3 Human Assessment Protocol

In this section, we introduce citation masking and recovery, a new approach for assessing the quality of evidence attribution.

Citation Masking Assume we have a claim c, a veracity label v, m evidence passages $E = \{e_1, e_2, e_3, ..., e_m\}$, and a generated explanation with n sentences $X = \{x_1, x_2, x_3, ..., x_n\}$.[3]

A subset of sentences $X_{cit} = \{x_i, x_{i+1}, ..., x_j\} \subseteq X$ contains inline citations (e.g. *[6]*),[4] where $1 \leq i \leq j \leq n$. We randomly select $e_k \in E$ $(1 \leq k \leq m)$ and mask its inline citation marker in the explanation (e.g. *you are wrong [6]* $\rightarrow$ *you are wrong*), producing X^{mask}, the masked explanation. We denote $X^{mask}_{cit} \subseteq X_{cit}$ as the subset of explanation sentences where their citation markers are removed.

Citation Recovery We next ask annotators to recover the masked sentences. That is, annotators are presented with claim c, veracity v, full evidence set E, evidence passage e_k, and masked explanation X^{mask}, and they are asked to find all sentences that should cite e_k. In other words, the task is to recover X^{mask}_{cit}. Denoting their prediction as X^{mask}_{pred}, a perfect identification means $X^{mask}_{pred} = X^{mask}_{cit}$. Note that X^{mask}_{cit} sometimes contains multiple sentences (e.g. when the explanation has 2 sentences that cite e_k) so this is a *multi-label* classification problem. To measure the degree of overlap between X^{mask}_{pred} and X^{mask}_{cit}, we compute set precision, recall, and F1. A high F1 performance indicates the explanation is accurate in its attribution and faithful to the source evidence.

[2]Detailed model version: GPT-4 = gpt-4-0613, GPT-3.5 = gpt-3.5-turbo, and LLaMA2 = Llama-2-7b-chat.

[3]Sentences are segmented with spaCy v3.7.2.

[4]Technically there is a possibility that $X = \emptyset$ when the LLM fails to cite any evidence passages, but we did not see this in practice.

Attribution score computation Given an annotator prediction X_{pred}^{mask} and the reference label set X_{cit}^{mask}, $Precision$ reflects the proportion of reference label in annotator prediction:

$$Precision = \frac{|X_{pred}^{mask} \cap X_{cit}^{mask}|}{|X_{pred}^{mask}|} \quad (1)$$

Similarly, $Recall$ reveals the proportion of reference label recovered by annotators:

$$Recall = \frac{|X_{pred}^{mask} \cap X_{cit}^{mask}|}{|X_{cit}^{mask}|} \quad (2)$$

$F1$ combines $Precision$ and $Recall$ via the harmonic mean.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

3.4 Human Annotation Details

Annotation Procedure We provide the annotator with the following information: a claim, veracity label, full evidence set, a selected evidence passage, and an explanation, and ask them to select (= highlight) sentences in the explanation that should cite the selected evidence passage (as in Figure 2). See Appendix F for the full annotation guidelines and interface.[5] Note that for each claim, we only select *one* random evidence passage to do citation recovery. For example in Figure 2, we select citation 9 to mask in the explanation. We address this limitation when we introduce LLM as annotator in Section 4, where *every* evidence passage is evaluated.

Annotator Recruitment The annotation task was done on Amazon Mechanical Turk (AMT).[6] We applied pre-screening pilot studies to find qualified annotators. We conducted individual reviews of submitted annotations, and offered feedback to annotators to address any misconceptions or confusion about the task. Annotators who performed well in the pilot study were invited to do a final round of annotation that produced the annotated data. In order to maintain high quality throughout the annotation, we used quality control questions to identify and remove poor-performing annotators. In total, we recruited 68 annotators for the final round of annotation.

[5]We also ask the annotators to rate the utility of the explanation after the citation recovery task, but we did not use these ratings for the experiments in this paper.

[6]https://www.mturk.com/

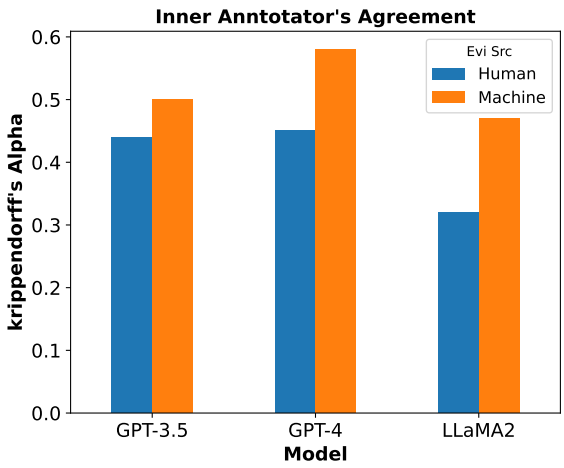

Figure 3: The inner-annotator agreement of human annotation. "Evi Src" indicates the whether the evidence is human or machine selected.

Quality Control Quality control was implemented for both the pilot study and final study. In the pilot study, each Human Intelligence Task (HIT) contained 3/6 control questions, 2 of which were positive questions containing exactly one answer each, and 1 of which is a negative question containing no correct answer (the original answer sentence has been removed). Annotators are expected to choose "There isn't any sentence that can correctly cite the highlighted core evidence." in such cases. All control questions were manually validated by the first author of the paper. An annotator who fails on any control questions is disqualified from participating in further tasks. In total, 9 pilot studies were released to recruit qualified annotators. For the final study, we use a different batch of control questions (i.e. there is no overlap in control questions between the pilot and final study). Workers with an accuracy of below 70% were removed and prevented from doing more tasks.

Annotator Compensation Annotators were paid USD$2.70 for each HIT. During the pilot study, workers were paid the base rate of $1 and a bonus of $1.70 if they passed the quality control. On average, a HIT took approximately 7 minutes to complete, resulting in a salary rate of $15 per hour. In the final study, all workers were compensated for their time, irrespective of whether they passed the quality control.

Annotator Agreement Each claim was annotated by 5 annotators. As agreement metrics like kappa (McHugh, 2012) are not applicable to our

Evi Src	Gen	Precision	Recall	F1
Human	GPT-4	**0.62±0.29**	**0.67±0.29**	**0.63±0.29**
	GPT-3.5	0.52±0.29	0.59±0.30	0.52±0.29
	LLaMA2	0.48±0.31	0.52±0.32	0.49±0.31
Machine	GPT-4	**0.72±0.32**	**0.79±0.31**	**0.74±0.31**
	GPT-3.5	0.55±0.39	0.55±0.39	0.53±0.37
	LLaMA2	0.49±0.39	0.51±0.40	0.49±0.38

Table 3: Human evaluation results of generated explanations. The "Evi Src" column indicates whether evidence is human- or machine-selected.

Retriever	Precision	Recall	F1
GPT-4	**0.40±0.21**	**0.75±0.23**	**0.47±0.20**
GPT-3.5	0.34±0.22	0.60±0.35	0.39±0.22
LLaMA2-70b	0.29±0.18	0.68±0.32	0.36±0.18

Table 4: Evidence retrieval performance.

multi-label scenario, we use Krippendorff's alpha (Krippendorff, 2011) to compute annotator agreement.[7] Figure 3 presents the inter-annotator agreement results. On average, GPT-4 explanations produce the best agreement, followed by GPT-3.5. LLaMA2 has the lowest agreement, because its generated explanations tend to be quite noisy (we will revisit this in the next section). Interestingly, explanations generated using human-selected evidence have a lower agreement score. Overall, these agreement results indicate moderate to good agreement for GPT-4 and GPT-3.5.

3.5 Human Evaluation Results

Given the human annotation, we now compute citation recovery performance (set precision, recall, and F1 in Section 3.3) to assess attribution quality in the generated explanations. Results are presented in Table 3.

GPT-4 has the best performance by a comfortable margin, suggesting that its generated explanations cite their sources more accurately than the other LLMs. Manual analysis reveals that LLaMA2 explanations often contain incorrect citations, such as *[1][2][3]* which are directly copied from the prompt/instruction, as well as random links. That said, even the best model (GPT-4 with machine-selected evidence) only produces an F1 of approximately 0.74, which means a good proportion of citations are still not faithful. There's also substantial variance (±0.31), indicating the performance is far worse in the worst case scenario.

Looking at human-selected vs. machine-selected evidence for a given LLM, we generally see similar attribution performance, implying that we may not need perfectly curated evidence for generating explanations. To understand how different the machine-selected evidence is to the human-selected evidence, we compute set precision, recall and F1 using the human-selected evidence as ground truth, and present the results in Table 4. Interestingly, across all LLMs we see only see a low to moderate F1 (0.3–0.5), meaning the machine-selected evidence is indeed quite different to human-selected evidence, but despite this difference, the resulting generated explanations do not differ much in terms of attribution quality.

4 Automatic Evaluation of Evidence Attribution

We now attempt to automate the annotation process, i.e. automate the citation recovery task in the assessment protocol (Section 3.3). We experiment with two approaches: Natural Language Inference (NLI), and LLM annotators.

NLI Most existing work frames the evaluation of citation quality as an NLI task (Gao et al., 2023). That is, given a pair of (premise, hypothesis), the NLI model outputs a label indicating whether the premise entails or contradicts the hypothesis. In our scenario, the premise is an evidence passage (e_k) while the hypothesis is an explanation sentence (a sentence in X). Given a selected evidence passage, we pair it with every sentence in the explanation and use a pretrained NLI model to detect instances of entailment: any sentence that is predicted to be entailed by the evidence passage is added to the recovered set (X_{pred}^{mask}).

In terms of pretrained NLI models, we use two models: DeBERTaV3 (He et al., 2023) and TRUE (Honovich et al., 2022).[8] The output label set for DeBERTaV3 is "entailment", "neutral" and "contradiction", and for TRUE is "1" (entailment) and "0" (not entailment). For our experiments we are only concerned with the entailment class.

LLM We also experiment with using an LLM to do the citation recovery task. We zero-shot prompt an LLM to recover the sentences (X_{pred}^{mask}), using an

[7] For Krippendorff's alpha, we use Jaccard distance to measure distance between two annotations (recall that each annotation result is a set, and as such we use Jaccard distance to measure the degree of set overlap).

[8] Detailed model versions: DeBERTaV3 = deberta-v3-base, TRUE = t5_xxl_true_nli_mixture.

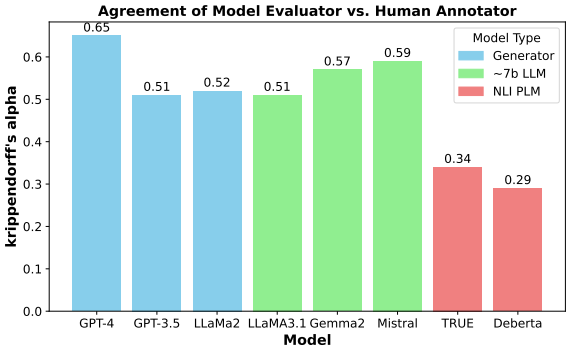

Figure 4: The agreement between automatic annotators vs. human annotators. "Generator" = models that are also used as (explanation) generators; "∼7b LLM" refers to open-source LLMs with around 7B parameters; "NLI PLM" indicates PLMs used to conduct pairwise NLI, as described in Section 4.

instruction similar to what we provide to the human annotators (Section 3.4). For the input, we include the evidence passage (e_k) and all explanation sentences in the prompt. See Appendix Table 13 for an example prompt.

In terms of LLMs, we use the same set of models that we used for generating explanations (i.e. GPT-4, GPT-3.5 and LLaMA2), and further include several smaller LLMs: LLaMA3.1 (Dubey et al., 2024), Mistral (Jiang et al., 2023) and Gemma (Mesnard et al., 2024).[9]

Evaluation Setting We perform automatic annotation under two settings: (1) *sampling*: for each claim, only one evidence passage is selected for masking – this is the approach we used in the human annotation in Section 3; and (2) *full*: for each claim, we mask and evaluate all evidence passages, one at a time. The *full* setting means we are assessing the attribution accuracy for every sentence that contains a citation marker in the explanation, and providing a more complete picture of the quality of the generated explanations (noting that this wasn't feasible with human evaluation). Note that in this setting, for each claim we have multiple set precision, recall and F1 scores (one for each evidence passage; Section 3.3), and we aggregate them by computing the mean.

[9]Detailed LLM versions: LLaMA3.1 = Llama-3.1-8B-Instruct, Mistral = Mistral-7B-Instruct-v0.3, Gemma = gemma-1.1-7b-it. All LLMs have at least 8K context window, and as such they fit the full prompt containing instructions, evidence set, and explanation.

4.1 Agreement with Human Annotation

To understand how reliable LLMs and NLI models are as automatic annotators, we first assess their agreement with human annotations. We compute the agreement between automated annotator and human annotator using krippendorff's alpha in a similar manner to that described in Section 3.5. But here we compute the agreement between two sets of annotation results: (1) an automatic annotator's results (e.g. GPT-4's); and (2) aggregate results from all human annotators.[10] We also standardise any labels outside the ground truth set (X_{cit}^{mask}) to a common label, as we are mostly interested in understanding the extent to which the automatic annotator selects the ground truth label when human annotators do so.[11]

We present the agreement results in Figure 4. We can see LLM annotators have a much higher correlation with human annotators than NLI annotators. Due to the nature of the NLI input, the hypothesis is an explanation sentence that is taken out of context of the whole explanation, and may not always be interpretable. The LLM also assesses each explanation sentence individually, but they do this by taking into account the whole explanation context (as it is provided as part of the prompt). Given these results, we focus on using LLMs as annotators for the remainder of our experiments.

4.2 Automatic Evaluation Results

We now present citation recovery results using LLMs as annotators for *sample* in Table 5 and *all* in Table 6. For the former we also include the human annotator results for comparison.

Comparing Automatic Annotator vs. Human Annotator In Table 5, we see that most automatic annotators have high agreement with human annotators: GPT-4 has the best attribution quality, followed by GPT-3.5 and LLaMA2. That said, if we look at the magnitude of the attribution quality, GPT-3.5 seems to overestimate this (i.e. its F1 numbers are much higher than those for humans), and this correlates with the agreement performance we saw in Section 4.1 where GPT-3.5 has a lower agreement. GPT-4 produces results that best align with human annnotators, and this is also reflected

[10]That is, we take the union of all recovered sentences (X_{pred}^{mask}) for each human annotator into a larger set.

[11]For example, if the ground truth set (X_{cit}^{mask}) is {9}, and the predicted set (X_{pred}^{mask}) from the automatic and human (grouped) annotator is {7, 9} and {5, 9} respectively, we will post-process both predicted sets into {-2, 9}.

Ann \ Gen	Evi Src	GPT-4	GPT-3.5	LLaMA2
		Attribution F1		
GPT-4	*Human*	0.69	0.60	0.50
	Machine	0.62	0.57	0.55
GPT-3.5	*Human*	0.83	0.73	0.61
	Machine	0.75	0.69	0.60
LLaMA2	*Human*	0.65	0.60	0.47
	Machine	0.57	0.51	0.47
LLaMA3.1	*Human*	0.57	0.49	0.45
	Machine	0.56	0.46	0.40
Gemma	*Human*	0.69	0.58	0.49
	Machine	0.59	0.52	0.46
Mistral	*Human*	0.58	0.49	0.42
	Machine	0.49	0.51	0.44
Human	*Human*	0.63	0.52	0.49
	Machine	0.74	0.53	0.49

Table 5: Evidence attribution results for the *sample* setting. The "Evi Src" column indicates whether evidence is *Human*- or *Machine*-selected, "Ann" the automatic annotators and "Gen" the explanation generators.

Ann \ Gen	Evi Src	GPT-4	GPT-3.5	LLaMA2
		Attribution F1		
GPT-4	*Human*	0.67	0.61	0.50
	Machine	0.60	0.50	0.40
GPT-3.5	*Human*	0.80	0.75	0.64
	Machine	0.74	0.57	0.45
LLaMA2	*Human*	0.59	0.58	0.48
	Machine	0.53	0.41	0.34
LLaMA3.1	*Human*	0.47	0.47	0.40
	Machine	0.44	0.36	0.26
Gemma	*Human*	0.67	0.63	0.50
	Machine	0.55	0.47	0.34
Mistral	*Human*	0.54	0.49	0.39
	Machine	0.48	0.40	0.30

Table 6: Evidence attribution results for the *full* setting.

in Section 4.1 where GPT-4 has the highest agreement.

Comparing *sample* vs. *full* Looking at both Table 5 and Table 6, we see a similar trend across both *sample* and *full*, where GPT-4 is consistently the best generator, followed by GPT-3.5 and LLaMA2. Interestingly though, if we look at the impact of evidence source ("human" vs. "machine"), for *sample* we see a small improvement from "machine" to "human", but this gap widens further in *full* (for generator GPT-3.5 and LLaMA2 in particular), suggesting that human-selected evidence does produce better explanations. This observation contrasts what we saw previously in Section 3.5 (where machine-selected evidence is on-par with human-selected evidence), and shows that it is important to assess attribution for every sentence in the explanation to get the complete picture.

Ann \ Gen	Evi Src	GPT-4	GPT-3.5	LLaMA2
		Transparency Proportion		
GPT-4	*Human*	31%	23%	17%
	Machine	9%	5%	4%

Table 7: The proportion of Explanation where all sentences are transparent. We use a threshold of 0.6 for attribution F1 score to classify instance as "transparent".

To provide a more intuitive understanding of the overall transparency of the generated explanations, we present Table 7 to show the proportion of explanations where every sentence attributes the source accurately (evaluated by GPT-4). We observe a large gap between "human" and "machine", again reinforcing that human-curated evidence leads to more transparent explanations.

Discussion Putting these results together, we can conclude as follows: (1) GPT-4 is the best model for generating fact-checking explanations; (2) there is still much room for improvement for explanation generation, because even with human-curated evidence, only 31% of GPT-4's explanations attribute the sources accurately; (3) GPT-4 is the best automatic annotator for assessing evidence attribution; and (4) the selection of evidence is crucial, and we found that human-selected evidence tends to produce higher quality explanations across different LLM generators.

5 Conclusion

In this paper, we explore evaluating evidence attribution for fact-checking explanations. We implement a novel protocol using both human annotation and automated methods, and show that our LLM-based annotations have stronger correlation with human annotations compared to NLI models. While the best-performing LLM shows promising results, there is still substantial room for improvement in explanation generation. Our findings highlight the importance of evidence selection, as human-curated evidence tends to produce more attributable explanations. Ultimately, our research provides insights into enhancing AI trustworthiness and paves the way for future developments of transparent explanation generation for fact-checking.

Limitations

This work only evaluates the evidence attribution of the generated explanations, i.e. how faithful they are based on the sources. We want to note that this evaluation does not assess factuality, which checks whether an explanation is factually correct, that said, there are already many studies on factuality, and our evaluation is orthogonal to these studies. Additionally, while we tried our best to craft the prompts, there remains the question of whether evidence attribution can be further improved with more prompt engineering. The PolitiHop dataset may contain out-of-date information, given it was developed a few years ago. As such, this may conflict with the LLM's parametric knowledge for certain claims. We saw this in our analyses where LLMs ignore evidence passages and generate statements based on their intrinsic knowledge. Also, the claims and evidence passages in PolitiHop are generally short single sentences. In practice, claims and evidence passages are likely to be longer.

Ethics Statement

All experiments were conducted under the approval of an internal ethic committee (ethics application ID: ANONYMISED). We provide workers with a plain language statement and consent form detailing the research goal, methods, compensation, and potential risks (e.g. misleading information). Workers are free to stop any time once they started and they will still be compensated.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Allison Baker and Viviane Fairbank. 2022. The truth in journalism fact-checking guide. https://thetijproject.ca/guide/how-to-fact-check/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6679–6692, Bangkok, Thailand. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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

Jie Huang and Kevin Chang. 2024. Citation: A key to building responsible and accountable large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 464–473, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock,

Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

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*, pages 14919–14935, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2018. Towards efficient machine translation evaluation by modelling annotators. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 77–82, Dunedin, New Zealand.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

T. Minka, J.M. Winn, J.P. Guiver, Y. Zaykov, D. Fabian, and J. Bronskill. 2018. /Infer.NET 0.3. Microsoft Research Cambridge. http://dotnet.github.io/infer.

Thomas P. Minka. 2001. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

OpenAI. 2024. Gpt-4 technical report.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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

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

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan

Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Rui Xing, Shraey Bhatia, Timothy Baldwin, and Jey Han Lau. 2022. Automatic explanation generation for climate science claims. In *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, pages 122–129, Adelaide, Australia. Australasian Language Technology Association.

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

A Assessing Perceived Utility

Although not the main focus of this paper, in addition to Evidence Attribution, we also assess Explanation Utility as a complementary metric. Utility captures whether users find an explanation helpful in clarifying the claim (Liu et al., 2023; Gao et al., 2023).

Utility Evaluation Utility measures to what extent the generated explanation clarifies a claim. Though a five-point Likert scale is commonly used (Liu et al., 2023; Gao et al., 2023), Ethayarajh and Jurafsky (2022) found that averaging using the Likert scale can result in a biased estimate. As such we use Direct Assessment (Graham et al., 2013) where annotators rate on a (continuous) sliding scale from 0–100, where 100 = best. We ask annotators the following question: *How helpful is the explanation in clarifying the truthfulness of the claim?*. We aggregate the judgements of multiple annotators for each explanation by computing the mean. We also applied a Bayesian model for utility score calibration, but it showed a similar tendency, so we decided to use the original scores.

Utility Score Calibration Perceiving the utility of fact-checking explanation is a subjective task. On the one hand, annotators might disagree on how useful the explanation is. Some workers may consistently provide low utility scores for all explanations due to their high standards, while others might be more lenient. Additionally, certain workers may only utilize a narrow range of the scoring scale such as the central part. On the other hand, while deploying our task on AMT provides a convenient and cost-effective solution, it comes with challenges such as high variance between workers, poor calibration, and the potential to draw misleading scientific conclusions (Karpinska et al., 2021).

Motivated by the aforementioned reasons, we used a simple Bayesian model (Mathur et al., 2018) to calibrate the annotated utility scores. The calibration functions as follows: assuming the utility score s is normally distributed around the true utility μ of the explanation, we use an accuracy parameter τ to model each worker's accuracy: higher value indicates smaller errors. The full generative modelling works as follows:

- For each explanation i, we draw true utility μ_i from the standard normal distribution.

$$\mu_i \sim \mathcal{N}(0, 1) \tag{4}$$

- Then for each annotator j, we draw their accuracy τ from a shared Gamma prior with shape parameter k and rate parameter θ [12].

$$\tau_j \sim \mathcal{G}(k, \theta) \tag{5}$$

- The annotator's utility score $s_{i,j}$ is then drawn from a normal distribution with mean μ_i and accuracy τ_j.

$$s_{i,j} \sim \mathcal{N}(\mu_i, \tau^{-1}) \tag{6}$$

Our goal is to maximize the likelihood of the observation of annotated utility score:

$$\begin{aligned} P(s) &= \int_j P(\tau_j) \int_i P(\mu_i) P(s_{i,j}|\mu_i, \tau) d\tau d\mu \\ &= \int_j \Gamma(\tau_j|k, \theta) \int_i \mathcal{N}(\mu_i|0, 1) \\ &\quad \mathcal{N}(s_{i,j}|\mu_j, \tau_j^{-1}) d\tau d\mu \end{aligned} \tag{7}$$

We first standardize individual annotators' utility scores via z-scoring to enhance comparability and reduce potential biases. Afterwards we use Expectation Propagation (Minka, 2001) to infer posterior over true utility score μ and annotator accuracy τ [13].

Utility Results High utility doesn't necessarily imply high evidence attribution. Although we found a a general pattern between utility and evidence attribution (Table 9), GPT-3.5 achieved the highest utility score when using human-selected evidence (72.47), even though its attribution score is much lower compared to GPT-4 (0.52 vs. 0.63). After filtering for transparent explanation with a threshold of $F1 = 1.0$ and $F1 \geq 0.9$, GPT-4 achieved the highest utility score, as shown in Table 10.This indicates that high utility GPT-3.5 explanations are not attributable. Overall, this demonstrates that utility and citation represent two distinct qualities and has an important implication: an explanation that appears helpful can actually still be misleading.

B Annotation Entropy

We used another metric Entropy (Shannon, 1948) to measure the randomness and the degree of uncertainty in the system:

$$H = -\sum p_k \, log \, p_k \tag{8}$$

[12] We use $k = 1.5$ and $\theta = 0.5$ based on our manual inspection of preliminary experiments.

[13] We implemented the model with Infer.NET framework (Minka et al., 2018).

Eval \ Gen	Evi Src	GPT-4 Precision	GPT-4 Recall	GPT-4 F1	GPT-3.5 Precision	GPT-3.5 Recall	GPT-3.5 F1	LLaMA2-70B Precision	LLaMA2-70B Recall	LLaMA2-70B F1
GPT-4	*Human*	0.59±0.32	0.96±0.18	0.69±0.25	0.50±0.32	0.88±0.30	0.60±0.29	0.40±0.33	0.76±0.43	0.50±0.34
	Machine	0.50±0.30	0.97±0.17	0.62±0.26	0.47±0.32	0.87±0.33	0.57±0.30	0.47±0.37	0.79±0.39	0.55±0.35
GPT-3.5	*Human*	0.82±0.34	0.88±0.32	0.83±0.32	0.72±0.39	0.79±0.38	0.73±0.37	0.59±0.45	0.68±0.47	0.61±0.45
	Machine	0.72±0.39	0.83±0.38	0.75±0.37	0.71±0.42	0.72±0.42	0.69±0.4	0.59±0.44	0.67±0.46	0.60±0.43
LLaMA2	*Human*	0.59±0.4	0.76±0.42	0.65±0.39	0.54±0.37	0.74±0.42	0.6±0.36	0.4±0.37	0.61±0.49	0.47±0.39
	Machine	0.49±0.35	0.75±0.43	0.57±0.35	0.48±0.41	0.62±0.47	0.51±0.4	0.41±0.37	0.61±0.48	0.47±0.39
LLaMA3.1	*Human*	0.47±0.31	0.80±0.40	0.57±0.32	0.41±0.33	0.70±0.45	0.49±0.34	0.36±0.31	0.66±0.47	0.45±0.35
	Machine	0.46±0.32	0.83±0.37	0.56±0.31	0.38±0.33	0.66±0.46	0.46±0.34	0.31±0.28	0.65±0.47	0.40±0.33
Gemma	*Human*	0.61±0.35	0.86±0.34	0.69±0.33	0.51±0.37	0.75±0.42	0.58±0.36	0.42±0.38	0.67±0.47	0.49±0.39
	Machine	0.52±0.38	0.78±0.41	0.59±0.37	0.47±0.39	0.68±0.45	0.52±0.38	0.42±0.41	0.59±0.48	0.46±0.40
Mistral	*Human*	0.46±0.27	0.90±0.29	0.58±0.26	0.38±0.28	0.82±0.37	0.49±0.28	0.32±0.25	0.76±0.42	0.42±0.28
	Machine	0.37±0.26	0.86±0.35	0.49±0.27	0.39±0.26	0.85±0.34	0.51±0.25	0.34±0.28	0.77±0.41	0.44±0.29
Human	*Human*	0.62±0.29	0.67±0.29	0.63±0.29	0.52±0.29	0.59±0.30	0.52±0.29	0.48±0.31	0.52±0.32	0.49±0.31
	Machine	0.72±0.32	0.79±0.31	0.74±0.31	0.55±0.39	0.55±0.39	0.53±0.37	0.49±0.39	0.51±0.40	0.49±0.38

Table 8: Full Evidence attribution results for the *sample* setting. The "Evi Src" column indicates whether evidence is *Human*- or *Machine*-selected, "Ann" the automatic annotators and "Gen" the explanation generators.

Evi Src	Gen	Utility	Entropy
Human	GPT-4	**0.28±0.18**	66.86±19.38
	GPT-3.5	0.41±0.18	**72.47±19.52**
	LLaMA2	0.34±0.16	65.13±18.00
Machine	GPT-4	**0.18±0.19**	**76.34±17.88**
	GPT-3.5	0.21±0.18	70.47±20.89
	LLaMA2	0.21±0.17	66.20±21.84

Table 9: Human evaluation results of Utility and Entropy in generated explanations. The "Evi Src" column indicates whether evidence is human- or machine-selected.

Evi Src	Gen	$F1 = 1.0$	$F1 \geq 0.9$
Human	GPT-4	**83.43±3.66**	**74.90±15.62**
	GPT-3.5	61.11±30.57	73.22±25.75
	LLaMA2	63.65±25.14	73.60±17.89
Machine	GPT-4	**84.09±7.94**	**83.88±7.86**
	GPT-3.5	83.62±6.74	83.15±6.75
	LLaMA2	82.96±6.90	82.96±6.90

Table 10: Utility score after filtering via $F1$ threshold, scores are presented in the form of ave±std

In multi-label scenario, the entropy of the label probability distribution reflects the likelihood of each chosen label. It also influences the probability of agreement on the label (Marchal et al., 2022). For instance, consider a certain annotation result $[0, 0, 4, 0]$, which represents the occurrences of each option, annotators exhibit high agreement in choosing the 2nd sentence to cite the reason (index starts from 0). In contrast, $[1, 1, 1, 1]$ shows evenly distributed choices on each option, which suggests greater uncertainty among annotators. Consequently, the latter will have higher entropy. In our task, we utilize entropy as an indicator of annotation uncertainty. We compute the normalized probability of each claim annotation and then apply equation 8 to calculate entropy.

C Prompts

We developed and optimized prompt for our task. We present evidence selection prompt in Table 14, explanation generation prompt in Table 12 and explanation evaluation prompt for LLM as annotators in Table 13.

D Experiment Details

We ran all offline models on 4 Nvidia A100 GPUs in a data parallel fashion. Explanation generation with GPT-4 and GPT-3.5 takes around 2 hours and LLaMA2 takes around 4 hours. The cost of generating explanations was $49 (US Dollar) with GPT-4 and $10 (US Dollar) with GPT-3.5. We present full evidence attribution results for *sample* setting in

E Scientific Artifacts

We list the licenses of different artifacts used in this paper: PolitiHop (MIT) [14], Huggingface Transformers (Apache License 2.0) [15], and spaCy (MIT) [16]. Our source code and annotated data will be under MIT license.

[14] https://github.com/copenlu/politihop
[15] https://github.com/huggingface/transformers
[16] https://github.com/explosion/spaCy

Eval \ Gen	Evi Src	GPT-4 Precision	GPT-4 Recall	GPT-4 F1	GPT-3.5 Precision	GPT-3.5 Recall	GPT-3.5 F1	LLaMA2-70B Precision	LLaMA2-70B Recall	LLaMA2-70B F1
GPT-4	*Human*	0.56±0.31	0.96±0.19	0.67±0.25	0.52±0.33	0.88±0.31	0.61±0.30	0.41±0.33	0.77±0.41	0.50±0.33
	Machine	0.50±0.32	0.95±0.21	0.60±0.28	0.41±0.33	0.76±0.41	0.50±0.32	0.33±0.35	0.61±0.48	0.40±0.36
GPT-3.5	*Human*	0.78±0.37	0.84±0.36	0.80±0.35	0.75±0.39	0.78±0.39	0.75±0.38	0.63±0.45	0.68±0.46	0.64±0.44
	Machine	0.70±0.39	0.82±0.38	0.74±0.37	0.58±0.46	0.61±0.46	0.57±0.44	0.44±0.46	0.50±0.49	0.45±0.46
LLaMA2	*Human*	0.53±0.39	0.71±0.45	0.59±0.4	0.53±0.39	0.68±0.45	0.58±0.39	0.43±0.4	0.59±0.49	0.48±0.41
	Machine	0.47±0.38	0.67±0.47	0.53±0.4	0.38±0.4	0.5±0.48	0.41±0.41	0.30±0.37	0.43±0.49	0.34±0.40
LLaMA3.1	*Human*	0.44±0.44	0.55±0.49	0.47±0.44	0.46±0.45	0.53±0.48	0.47±0.44	0.38±0.44	0.46±0.49	0.40±0.44
	Machine	0.40±0.43	0.52±0.50	0.44±0.44	0.35±0.44	0.40±0.47	0.36±0.43	0.25±0.40	0.29±0.45	0.26±0.40
Gemma	*Human*	0.60±0.37	0.83±0.37	0.67±0.35	0.57±0.38	0.79±0.39	0.63±0.36	0.45±0.40	0.65±0.47	0.50±0.40
	Machine	0.48±0.39	0.74±0.44	0.55±0.39	0.44±0.41	0.60±0.47	0.47±0.40	0.31±0.41	0.44±0.49	0.34±0.41
Mistral	*Human*	0.41±0.26	0.89±0.31	0.54±0.26	0.37±0.26	0.81±0.37	0.49±0.28	0.29±0.26	0.69±0.45	0.39±0.29
	Machine	0.37±0.27	0.85±0.36	0.48±0.28	0.31±0.29	0.69±0.45	0.40±0.31	0.22±0.26	0.57±0.49	0.30±0.30

Table 11: Full Evidence attribution results for the *full* setting. The "Evi Src" column indicates whether evidence is *Human*- or *Machine*-selected, "Ann" the automatic annotators and "Gen" the explanation generators.

Prompt
Instructions: You are required to write an accurate, coherent and logically consistent explanation for the claim based on the given veracity and list of reasons in one paragraph. Use an unbiased and journalistic tone. When citing several search results, use [1][2][3]. Ensure that each reason is cited only once. Do not cite multiple reasons in a single sentence. Reasons: Reason [1] What's more, the picture referenced in the Facebook post alleging that anglerfish are typically 7 feet is taken from the Australian Museum's 2012 exhibit titled "Deep Oceans. Reason [2] When the exhibit opened in June 2012, The Sydney Morning Herald reported on how the exhibit's team had created an "oversized anglerfish" and listed the many steps in making it: "Pieces such as the oversized anglerfish, with huge fangs and antenna-like flashing rod to attract prey, begin with cutting and welding a metal frame, then sculpting material over it and, finally, hand painting it," the story says. Claim:The typical anglerfish is seven feet long. Veracity: False Explanation:

Table 12: Prompts for explanation generation

Prompt
Find the most suitable explanation sentence(s) that can cite the given reason sentence. Return the sentence number(s) separated by comma, e.g., 0 or 0,2. Return -1 if no suitable sentences are found. Consider semantic citation relationships, not just keyword matching. Only return numbers, DO NOT include any additional output. Reason Sentence: What's more, the picture referenced in the Facebook post alleging that anglerfish are typically 7 feet is taken from the Australian Museum's 2012 exhibit titled "Deep Oceans. Explanation Sentences: 1. The claim that the typical anglerfish is seven feet long is false. 2. One reason for this is that the picture used in the Facebook post to support the claim is taken from the 2012 exhibit titled D̈eep Oceansät the Australian Museum, which featured an översized anglerfisḣ[8]. 3. This suggests that the anglerfish showcased in the exhibit is not typical in size.", 4. Furthermore, an article from The Sydney Morning Herald that reported on the exhibit mentioned the many steps involved in creating the exhibit's översized anglerfisḣ[10]." 5. These findings indicate that the anglerfish portrayed in the picture and the exhibit are not representative of the typical size of anglerfish. Answers:

Table 13: Prompts for automatic explanation evaluation

Prompt
Instructions: You are required to retrieve a subset of reasons from the provided full reasons. The sentences in this subset should be coherent and logically consistent, presenting the most crucial information necessary to establish the veracity of the claim. Aim for the minimum number of sentences in the subset while maintaining the completeness and clarity. When extract reasons, use [1,2,3]. At last, provide a justification explaining why they are good reasons and how they form a logically consistent reasoning process. Demonstration: Reasons: Reason [0]: Anglerfish may have a reputation for being among the creepier-looking ocean-dwellers, but it's not because they grow to be seven feet long, as a viral image on Facebook claims. Reason [1]: The Jan. 12 post shows a young girl reaching toward what appears to be a very large anglerfish mounted on display at a museum. Reason [2]: The text above the image reads, "So,... I've spent my entire life thinking the Deep Sea Angler Fish was about the size of a Nerf football. Reason [3]: What's more, the picture referenced in the Facebook post alleging that anglerfish are typically 7 feet is taken from the Australian Museum's 2012 exhibit titled "Deep Oceans". Reason [4]: The anglerfish in the photo is actually a large-scale sculpture model of the fish made of plaster. Reason [5]: When the exhibit opened in June 2012, The Sydney Morning Herald reported on how the exhibit's team had created an "oversized anglerfish" and listed the many steps in making it: "Pieces such as the oversized anglerfish, with huge fangs and antenna-like flashing rod to attract prey, begin with cutting and welding a metal frame, then sculpting material over it and, finally, hand painting it," the story says. Claim: The typical anglerfish is seven feet long. Veracity: False Extracted Reasons: [3,5] Justification: Reason [3] establishes that the Facebook post's claim relies on a picture from the Australian Museum's 2012 exhibit. Reason [5] then reveals that the anglerfish in the exhibit is an oversized sculpture, not an actual specimen. Together, these reasons logically demonstrate that the viral claim of typical anglerfish being seven feet long is false, as it is based on a misrepresented image from an exhibit. Here's the actual task: Reasons: Reason [0]: Amid fears about the coronavirus disease, a YouTube video offers a novel way to inoculate yourself: convert to Islam. Reason [1]: "20m Chinese gets converted to Islam after it is proven that corona virus did not affect the Muslims," reads the title of a video posted online Feb. 18. Reason [2]: The footage shows a room full of men raising an index finger and reciting what sounds like the Shahadah, a statement of faith in Islam. Reason [3]: That's because the footage is from at least as far back as May 26, 2019, when it was posted on Facebook with this caption: "Alhamdulillah welcome to our brothers in faith." Reason [4]: On Nov. 7, 2019, it was posted on YouTube with this title: "MashaaAllah hundreds converted to Islam in Philippines." Reason [5]: Both posts appeared online before the current outbreak of the new coronavirus, COVID-19, was first reported in Wuhan, China, on Dec. 31, 2019. Reason [6]: But even if the footage followed the outbreak, Muslims are not immune to COVID-19, as the Facebook post claims. Reason [7]: After China, Iran has emerged as the second focal point for the spread of COVID-19, the New York Times reported on Feb. 24. Reason [8]: "The Middle East is in many ways the perfect place to spawn a pandemic, experts say, with the constant circulation of both Muslim pilgrims and itinerant workers who might carry the virus." Reason [9]: On Feb. 18, Newsweek reported that coronavirus "poses a serious risk to millions of inmates in China's Muslim prison camps." Claim: It was stated on February 18, 2020 in a YouTube post: "20 million Chinese converted to Islam after it's proven that the coronavirus doesn't affect Muslims." Veracity: False Extracted Reasons:

Table 14: Prompts for evidence selection.

F Annotation Interface

Determine the transparency and helpfulness of explanations for fact checking claims

A **claim** is a statement or assertion that declares something to be true, and it can be either truthful or bogus. In the fact-checking process, an **explanation** is essential. Your task is to assess whether a provided explanation is transparent and helpful, specifically in terms of **appropriately citing relevant sources**.

In this annotation, you will be presented 2 claims. For each claim, you'll be given a list of supporting/debunking reasons and a succinct explanation that clarifies the claim based on the listed reasons.

For each claim, you'll be asked to perform 2 tasks: (1) find missing citation sentences in the explanation; and (2) judge the helpfulness or utility of the explanation in clarifying the claim.

In the upcoming section, we'll begin by providing two illustrative **examples** of the annotation task. This is intended to assist you in grasping the idea of the task. **Please read the example carefully.** Once you click "Proceed" at the bottom of the page, you will move on to the primary annotation and encounter the first actual claim.

Important information: Thank you so much for participating in the task. Please kindly read the example carefully and complete the task. To ensure quality, we will manually review your responses. Bonuses will be provided only for careful task completion. Please feel free to reach out if you have any questions or suggestions. Each worker please kindly complete no more than 5 HITs in this round.

Annotation Example

Example Claim 1 (Task1)

Please read the instructions carefully before proceeding.

Instructions

In the first task, you'll be presented a claim, core reasons, a conclusion about the truthfulness of the claim, and a succinct explanation that cites these core reasons. In the explanation, you'll see square brackets citations that refer to core reasons (e.g. *highlighting the conversion of hundreds of individuals to Islam in the Philippines*[5]). In this task, **one core reason is highlighted**, its corresponding citation mark are **missing** in explanation sentences that cite this reason. Your task is to identify these sentences by clicking on them. Click on it again will cancel selection. Please note there might be **zero, one or multiple** sentences that cite the highlighted reason. If you cannot find any suitable sentence, please click the button that says "There isn't any sentence that can cite the highlighted core reason".

Claim

It was stated on February 18, 2020 in a YouTube post: “20 million Chinese converted to Islam after it’s proven that the coronavirus doesn’t affect Muslims.”

Full list of reasons

Figure 5: Annotation interface Page 1

Determine the transparency and helpfulness of explanations for fact checking claims

A **claim** is a statement or assertion that declares something to be true, and it can be either truthful or bogus. In the fact-checking process, an **explanation** is essential. Your task is to assess whether a provided explanation is transparent and helpful, specifically in terms of **appropriately citing relevant sources**.

In this annotation, you will be presented 2 claims. For each claim, you'll be given a list of supporting/debunking reasons and a succinct explanation that clarifies the claim based on the listed reasons.

For each claim, you'll be asked to perform 2 tasks: (1) find missing citation sentences in the explanation; and (2) judge the helpfulness or utility of the explanation in clarifying the claim.

In the upcoming section, we'll begin by providing two illustrative **examples** of the annotation task. This is intended to assist you in grasping the idea of the task. **Please read the example carefully.** Once you click "Proceed" at the bottom of the page, you will move on to the primary annotation and encounter the first actual claim.

Important information: Thank you so much for participating in the task. Please kindly read the example carefully and complete the task. To ensure quality, we will manually review your responses. Bonuses will be provided only for careful task completion. Please feel free to reach out if you have any questions or suggestions. Each worker please kindly complete no more than 5 HITs in this round.

Annotation Example

Example Claim 1 (Task1)

Please read the instructions carefully before proceeding.

Instructions

In the first task, you'll be presented a claim, core reasons, a conclusion about the truthfulness of the claim, and a succinct explanation that cites these core reasons. In the explanation, you'll see square brackets citations that refer to core reasons (e.g. *highlighting the conversion of hundreds of individuals to Islam in the Philippines*[5]). In this task, **one core reason is highlighted**, its corresponding citation mark are **missing** in explanation sentences that cite this reason. Your task is to identify these sentences by clicking on them. Click on it again will cancel selection. Please note there might be **zero, one or multiple** sentences that cite the highlighted reason. If you cannot find any suitable sentence, please click the button that says "There isn't any sentence that can cite the highlighted core reason".

Claim

It was stated on February 18, 2020 in a YouTube post: “20 million Chinese converted to Islam after it’s proven that the coronavirus doesn’t affect Muslims.”

Full list of reasons

Figure 6: Annotation interface Page 2

If you are interested in the context of core reasons, we provide a full list of reasons where core reasons are extracted. You can view this list by clicking the "Expand" button.

1. Amid fears about the coronavirus disease, a YouTube video offers a novel way to inoculate yourself: convert to Islam.
2. "20m Chinese gets converted to Islam after it is proven that corona virus did not affect the Muslims," reads the title of a video posted online Feb. 18 .
3. The footage shows a room full of men raising an index finger and reciting what sounds like the Shahadah, a statement of faith in Islam.
4. That's because the footage is from at least as far back as May 26, 2019, when it was posted on Facebook with this caption: "Alhamdulillah welcome to our brothers in faith."
5. On Nov. 7, 2019, it was posted on YouTube with this title: "MashaaAllah hundreds converted to Islam in Philippines."
6. Both posts appeared online before the current outbreak of the new coronavirus, COVID-19, was first reported in Wuhan, China, on Dec. 31, 2019.
7. But even if the footage followed the outbreak, Muslims are not immune to COVID-19, as the Facebook post claims.
8. After China, Iran has emerged as the second focal point for the spread of COVID-19, the New York Times reported on Feb. 24 .
9. "The Middle East is in many ways the perfect place to spawn a pandemic, experts say, with the constant circulation of both Muslim pilgrims and itinerant workers who might carry the virus."
10. On Feb. 18, Newsweek reported that coronavirus "poses a serious risk to millions of inmates in China's Muslim prison camps."

Collapse

Core reasons

- 4. That's because the footage is from at least as far back as May 26, 2019, when it was posted on Facebook with this caption: "Alhamdulillah welcome to our brothers in faith."
- 5. On Nov. 7, 2019, it was posted on YouTube with this title: "MashaaAllah hundreds converted to Islam in Philippines."
- **6. Both posts appeared online before the current outbreak of the new coronavirus, COVID-19, was first reported in Wuhan, China, on Dec. 31, 2019.**

Conclusion

TRUE

Explanation

The claim that 20 million Chinese converted to Islam after the coronavirus was proven to not affect Muslims is false. The viral footage that is being used as evidence is actually older than the current pandemic. The same video was published on Facebook on May 26, 2019, with a caption that welcomed new brothers in faith [4]. Moreover, the video made its way to YouTube on November 7, 2019, highlighting the conversion of hundreds of individuals to Islam in the Philippines [5]. These dates clearly predate the first reporting of the new coronavirus (COVID-19) outbreak in Wuhan, China, on December 31, 2019. Subsequently, the claim's timeline is inconsistent with the established chronology of these two events in Nov and Dec [5]. Therefore, the assertion in the claim regarding mass conversions due to the

Figure 7: Annotation interface Page 3

coronavirus is not corroborated by the timeline of events revealed in the sources of the video.

○ There isn't any sentence that can correctly cite the highlighted core reason.

Example Answer (Click to see how to annotate)

- These dates clearly predate the first reporting of the new coronavirus (COVID-19) outbreak in Wuhan, China, on December 31, 2019 [6].
- Subsequently, the claim's timeline is inconsistent with the established chronology of these two events in Nov and Dec [5][6].

  These two sentences accurately represent the information presented in the cited sentence. It maintains faithfulness to the original statement by conveying the same information about the timing of the posts and the location of COVID-19 outbreak (Wuhan). It helps the reader easily trace the source of information back to its origin, promoting transparency and credibility in the text.

Continue

(once clicked you will move to the next task and can't return)

Example Claim 1 (Task2)

Please read the instructions carefully before proceeding.

Instructions

In the second task, you'll be presented the same claim, explanation and conclusion as before, and your job is to rate how helpful is the explanation in clarifying the truthfulness of the claim.

Claim

It was stated on February 18, 2020 in a YouTube post: “Says 20 million Chinese converted to Islam after it’s proven that the coronavirus doesn’t affect Muslims.”

Explanation

The claim that 20 million Chinese converted to Islam after the coronavirus was proven to not affect Muslims is false. The viral footage that is being used as evidence is actually older than the current pandemic. The same video was published on Facebook on May 26, 2019, with a caption that welcomed new brothers in faith [4]. Moreover, the video made its way to YouTube on November 7, 2019, highlighting the conversion of hundreds of individuals to Islam in the Philippines [5]. These dates clearly predate the first reporting of the new coronavirus (COVID-19) outbreak in Wuhan, China, on December 31, 2019 [6]. Therefore, the assertion in the claim regarding mass conversions due to the coronavirus is not corroborated by the timeline of events revealed in the sources of the video.

Conclusion

TRUE

Rating

Figure 8: Annotation interface Page 4

How helpful is the explanation in clarifying the truthfulness of the claim?

Less helpful — More helpful

0% 25% 50% 75% 100%

[Please select the value using the sliding bar.]

Example Answer

- Example rating: 98%

  The explanation is very helpful in clarifying the truthfulness of the claim. It provides a clear timeline of events that shows the claim is false. The generated explanation are coherent, fluent and the citation is faithful.

Continue

(once clicked you will move to the next task and can't return)

Example Claim 2 (Task1)

Claim

The typical anglerfish is seven feet long.

Full list of reasons

If you are interested in the context of core reasons, we provide a full list of reasons where core reasons are extracted. You can view this list by clicking the "Expand" button.

Expand

Core reasons

- **8. What's more, the picture referenced in the Facebook post alleging that anglerfish are typically 7 feet is taken from the Australian Museum's 2012 exhibit titled "Deep Oceans."**
- 10. When the exhibit opened in June 2012, The Sydney Morning Herald reported on how the exhibit's team had created an "oversized anglerfish" and listed the many steps in making it: "Pieces such as the oversized anglerfish, with huge fangs and antenna-like flashing rod to attract prey, begin with cutting and welding a metal frame, then sculpting material over it and, finally, hand painting it," the story says.

Conclusion

TRUE

Explanation

Figure 9: Annotation interface Page 5

The claim that the typical anglerfish is seven feet long is false. Information sourced from a 2021 article on marine life reveals that an image featured in a Facebook post, purporting to depict an accurate representation of a common anglerfish, is accurately presented as an authentic specimen. At the time of the exhibit's opening, The Sydney Morning Herald noted that the display contained an "oversized anglerfish," which was fabricated using sculpting materials and a metal frame [10]. Thus, it's clear that the claim is based on an artificial construct not representing a typical anglerfish's size.

○ There isn't any sentence that can cite the highlighted core reason.

Example Answer

- There isn't any sentence that can correctly cite the highlighted core reason.

  Though the second sentence seems plausible and relevant to the highlighted core reason, it is not a suitable citation. It misrepresents the information and inaccurately claims that the Facebook post's image is accurately presented as an authentic specimen, while the highlighted core reason clearly states that the image is sourced from an exhibit, suggesting it's not authentic. Incorrect year 2021 in the explanation, it should be 2012. This misrepresentation in this generated explanation compromises faithfulness, factuality.

Continue

(once clicked you will move to the next task and can't return)

Example Claim 2 (Task2)

Please read the instructions carefully before proceeding.

Instructions

In the second task, you'll be presented the same claim, explanation and conclusion as before, and your job is to rate how helpful is the explanation in clarifying the truthfulness of the claim.

Claim

The typical anglerfish is seven feet long.

Explanation

The claim that the typical anglerfish is seven feet long is false. Information sourced from a 2021 article on marine life reveals that an image featured in a Facebook post, purporting to depict an accurate representation of a common anglerfish, is accurately presented as an authentic specimen [8]. At the time of the exhibit's opening, The Sydney Morning Herald noted that the display contained an "oversized anglerfish," which was fabricated using sculpting materials and a metal frame [10]. Thus, it's clear that the claim is based on an artificial construct not representing a typical anglerfish's size.

Conclusion

TRUE

Figure 10: Annotation interface Page 6

Rating

How helpful is the explanation in clarifying the truthfulness of the claim?

Less helpful — More helpful

0% 25% 50% 75% 100%

[Please select the value using the sliding bar.]

Example Answer

- Example rating: 10%

  While the generated explanation is fluent, it is not helpful in clarifying the truthfulness of the claim. The citation is not suitable. It misrepresents important information in the core reasons and caused confusion.

☑ I have read the Plain Language Statement and Consent Form and agree to work on the task.

Proceed

Figure 11: Annotation interface Page 7