Evaluating Transparency of Machine Generated Fact Checking Explanations

Anonymous ACL submission

Abstract

An important factor when it comes to generating fact-checking explanations is the *selection of evidence*: intuitively, high-quality explanations can only be generated given the right evidence. In this work, we investigate the impact of human-curated vs. machine-selected evidence for explanation generation using large language models. To assess the quality of explanations, we focus on transparency (whether an explanation cites sources properly) and utility (whether an explanation is helpful in clarifying a claim). Surprisingly, we found that large language models generate similar or higher quality explanations using machine-selected evidence, suggesting carefully curated evidence (by humans) may not be necessary. That said, even with the best model, the generated explanations are not always faithful to the sources, suggesting further room for improvement in explanation generation for fact-checking. Code and data are available here: ANONYMISED.

1 Introduction

Automated fact-checking systems are designed to classify claims based on their truthfulness (Guo et al., 2022). However, debunking via simply calling it "false" can trigger a backfire effect, where the belief of false claims is further reinforced rather than hindered, reducing their effectiveness (Lewandowsky et al., 2012). As such, there's growing research on generating textual explanations to justify the outcomes of fact-checking systems (Guo et al., 2022; Russo et al., 2023).

Large language models (LLMs) have been used to generate explanations in various contexts (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 into the LLM for explanation generation. We present Figure 1 to illustrate the fact-checking explanation generation process.

How to perform effective *evidence selection* is an important question. Intuitively, high-quality explanations can only be generated given the right evidence. This constitutes the core research question of this paper: is carefully hand-curated evidence needed for explanation generation, or does machine-selected evidence suffice?

To answer this question, we need to first define how we evaluate the quality of explanations. We focus on two traits: transparency and utility. Transparency assesses whether an explanation cites the sources evidence appropriately (Gao et al., 2023); this is motivated by journalistic practices in fact-checking, where experts carefully track their sources to ensure every cited statement is precise (Baker and Fairbank, 2022). Utility, on the other hand, captures whether users find an explanation helpful in clarifying the claim (Liu et al., 2023). We evaluate transparency and utility manually. For transparency, we introduce a new approach where we mask citations in explanations and ask humans to recover them: the idea is that explanations that cite the sources correctly should produce a higher identification accuracy.

Our findings are: (1) machine-selected evidence is either on-par or superior to manually-selected evidence for explanation generation, depending on the LLM; (2) machine-selected evidence, based on manual analysis, appears to include a larger set of evidence passages that are relevant to the claim; and (3) despite encouraging performance using LLMs for generating fact-checking explanations, there is still room for improvement.

2 Automatic Explanation Generation

Our LLM-based explanation generation system assumes the following as input: (1) a claim; (2) a (large) set of evidence passages, some of which may be useful for debunking the claim; and (3)

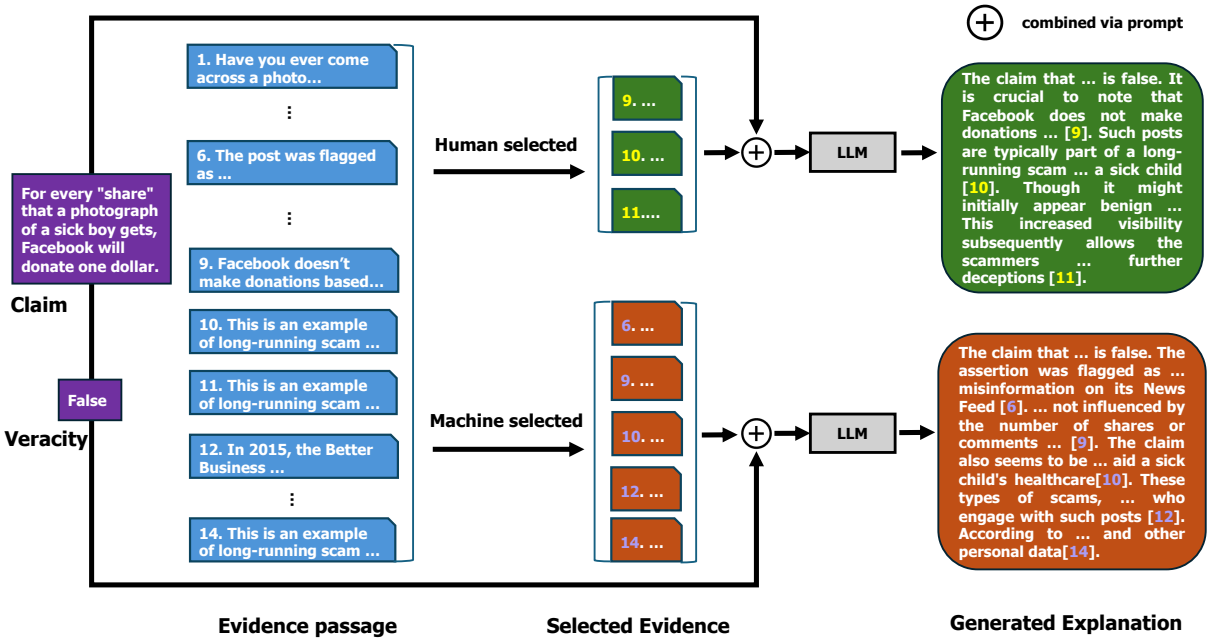


Figure 1: Automated explanation generation for fact-checking. Given a claim and a list of evidence passages, a subset of these passages is selected, either by humans or machines, and input into a large language model (LLM) along with the claim to generate the explanation.

the veracity of the claim (e.g. *True*).[1] 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 whole process.

Evidence Selection We experiment with two methods: (1) (one-shot) prompt an LLM (Appendix Table 6); or (2) ask human judges to select a subset of the most relevant evidence passages given a claim and the full set of evidence passages.[2]

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 (Appendix Table 5); 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). We test a range of LLMs in our experiments, noting that we always use the *same* LLM for evidence selection and explanation generation.

3 Assessing Explanation Transparency and Perceived Utility

We use Amazon Mechanical Turk to source workers to evaluate two qualities of the generated explanations: **transparency**, which assesses whether an explanation is citing the evidence passages appropriately; and **utility**, which measures the extent to which an explanation helps clarify the claim.

Transparency Evaluation 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]*), 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 without citation markers.

We next ask annotators to recover the masked sentences (see Appendix D for more annotation details). That is, annotators are presented with claim

[1]The rationale for assuming the veracity label as input is we see our explanation generation system being applied to a fact-checking classification system to provide a layer of interpretability in a real-world application.

[2]The dataset we use for our experiments contains the human-selected evidence passages, so we manually source this ourselves (see Section A).

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

Setting	Model	Precision	Recall	F1	Entropy	Utility
Human	GPT4	**0.62±0.29**	**0.67±0.29**	**0.63±0.29**	**0.28±0.18**	66.86±19.38
	GPT35	0.52±0.29	0.59±0.30	0.52±0.29	0.41±0.18	**72.47±19.52**
	LLaMA2-70b	0.48±0.31	0.52±0.32	0.49±0.31	0.34±0.16	65.13±18.00
Machine	GPT4	**0.72±0.32**	**0.79±0.31**	**0.74±0.31**	**0.18±0.19**	**76.34±17.88**
	GPT35	0.55±0.39	0.55±0.39	0.53±0.37	0.21±0.18	70.47±20.89
	LLaMA2-70b	0.49±0.39	0.51±0.40	0.49±0.38	0.21±0.17	66.20±21.84

Table 1: Evaluation results of generated explanations for transparency ("Precision", "Recall", "F1" and "Entropy") and utility ("Utility"). The "Setting" column indicates whether evidence is human- or machine-selected.

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 not a trivial judgement. To measure the degree of overlap between X^{mask}_{pred} and X^{mask}_{cit}, we compute set precision, recall, and F1 (see Appendix B).

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.[4] We aggregate the judgements of multiple annotators for each explanation by computing the mean.[5]

4 Experiments

Dataset We use PolitiHop (Ostrowski et al., 2021) as our dataset for all experiments. An instance contains a claim, a veracity label, a set of evidence passages, and a subset of human-selected evidence passages. The human-selected evidence passages are a set of relevant passages that clarify the claim. For each claim, PolitiHop contains multiple human-selected evidence sets; we randomly select one for our experiments.

Large Language Models We experiment with three language models: GPT-4 (OpenAI, 2024) (gpt-4-0613, accessed Aug-Nov, 2023), GPT-3.5 Turbo (gpt-3.5-turbo-16k, accessed Aug-Nov, 2023) and LLaMA2-70B (Touvron et al., 2023).[6] 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. We generate explanations for 100 claims sampled from PolitiHop. See the Table 4 in the Appendix for generation statistics.

Evaluation In addition to citation precision/recall/F1 for transparency and mean utility scores (Appendix B), we also compute the entropy of the distribution of human judgements for transparency.[7] Entropy tells us how well the annotators agree with each other: a low entropy means annotators selected a similar set of sentences.

5 Results and Discussion

Though competitive, there's still room for LLMs to improve on transparency. As shown in Table 1, GPT-4 is the most transparent model (in terms of precision, recall, F1, and Entropy) by a comfortable margin, followed by GPT-35 and LLaMA2. 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, limiting the real-world applicability.

[4]We ask annotators the following question: *How helpful is the explanation in clarifying the truthfulness of the claim?*

[5]We also applied a Bayesian model for utility score calibration, but it showed a similar tendency, so we decided to use the original scores. See Appendix F for details.

[6]We also tested other models (LLaMA2-7B, FlanT5-xxl (Chung et al., 2024), Falcon-30B (Almazrouei et al., 2023) and MPT (Team, 2023)) but excluded them because these models generated repeated content with fabricated citations.

[7]Concretely, an outcome corresponds to an explanation sentence, and the weight for an outcome (sentence) is the fraction of annotators who selected it.

Machine-selected evidences result in similar or better transparency and utility scores for explanations. Table 1 shows that for GPT-4, machine-selected evidence produces substantially better explanations in terms of transparency (F1 = 0.74 vs. 0.63) and utility (76.34 vs. 66.86). For GPT-35 and LLaMA2-70b, the results are similar,[8] suggesting that machine-selected evidence is at least as good as that selected by humans.

Machine-selected evidence is more comprehensive. Our previous finding suggests two possibilities: (1) LLMs are tolerant to noisy evidence for explanation generation; or (2) machine-selected evidence is higher-quality in the first place. Looking at the evidence retrieval performance (Table 2) where we measure how well the machine-selected evidence set overlaps with the human-selected evidence set, we see that GPT-4 performs best compared to other LLMs, but in absolute terms (F1 = 0.47) the machine-selected set is still substantially different to the human-selected set. To understand the quality of machine-selected evidence better, we manually examine the GPT-4-selected evidence and found that GPT-4 selected more evidence passages on average (5 vs. 3 per claim) and as such the evidence set is more comprehensive. We also notice that there is sometimes redundancy in GPT-4 selected evidence passages, but this isn't necessarily a negative outcome as it can further reinforce a key point in the explanation.

High utility doesn't necessarily imply high transparency. Although we found a a general correlation between utility and transparency (Table 1), GPT-35 achieved the highest utility score when using human-selected evidence (72.47), even though its transparency is much lower compared to GPT-4 (0.52 vs. 0.63). 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.

Transparency is hindered by LLM's parametric knowledge. We manually analyse instances with the lowest F1 (transparency) and found that the majority of them (60%) include fabricated statements, likely generated using the LLM's parametric knowledge and not the evidence passages.

[8]The exception here is Entropy, where we see a much lower entropy for machine-selected evidence (0.21 vs. 0.41 for GPT-35; 0.21 vs. 0.34 for LLaMA2-70b); this result implies machine-selected evidence is somewhat better (as it leads to higher annotator agreement for recovering the masked sentences).

Model	Precision	Recall	F1
GPT-4	**0.40±0.21**	**0.75±0.23**	**0.47±0.20**
GPT-35	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 2: Evidence retrieval performance.

6 Related Work

Automated fact-checking aims to classify the veracity of a claim (Guo et al., 2022; Russo et al., 2023). However, debunking by giving a classification label (e.g. *False*) is not persuasive 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). But extractive methods struggle to produce explanations with sufficient context and abstractive methods are prone to hallucination (Russo et al., 2023). To address hallucination, explanation generation models now incorporate citations to sources (Huang and Chang, 2023). Liu et al. (2023) introduced the evaluation of citation quality and used human judges to audit the verifiability of popular commercial generative engines. Gao et al. (2023) developed automatic metrics along three dimensions: fluency, correctness, and citation quality, to assesss LLM for question answering.

7 Conclusion

We explore the impact of human vs. machine-selected evidence for generating fact-checking explanations with LLMs. Surprisingly, we found that machine-selected evidence is either on-par or superior to human-selected evidence, depending on the LLM. Further analyses reveal that machine-selected evidence is more comprehensive and tends to contain redundant evidence (which help reinforce a key point). Our results also show that an explanation that appear helpful can be misleading in terms of how it cites the sources, highlighting the need to assess utility and transparency separately. Lastly, even though the best LLM exhibits encouraging performance, there's still much room for improvement for explanation generation.

Limitation

This work only evaluates the transparency of the generated explanations, i.e. how faithful they are based on the sources. We want to note that this transparency evaluation does not assess factuality, which checks whether an explanation is factually correct. Additionally, while we tried our best to craft the prompts, there remains the question of whether transparency 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

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

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/.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

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

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.

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.

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.

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.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

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.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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.

Claim: Shands Hospital of Gainesville Florida has confirmed its first case of coronavirus. **passage 1:** Anglerfish may have a reputation for being among the creepier-looking ocean-dwellers.. ... **passage 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". ... **passage 10 :** When the exhibit opened in June 2012, The Sydney Morning Herald reported on how the exhibit2019s team had created an "oversized anglerfish"...
Veracity: False
Evidences: {8,10}
Generated Explanation: The claim that the typical anglerfish is seven feet long is false. The confusion originates from a picture published in a 2012 Facebook post, which was actually taken from the Australian Museum's exhibit "Deep Oceans"[8]. When this exhibit was inaugurated in the same year, The Sydney Morning Herald revealed that the exhibit's team had developed an "oversized anglerfish", indicating that the size of the anglerfish in the picture is not representative of the typical size for the species [10].

Table 3: An example based on PolitiHop with claim c, veracity v, associated evidences E. X is the generated explanation that cites E.

A PolitiHop Dataset Sample

Table 3 shows an example from PolitiHop. As illustrated in Figure 1, LLMs produce "Generated Explanation" that cite "Evidences".

B Transparency Evaluation Details

Given annotators' 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 harmonic mean.

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

Entropy (Shannon, 1948) is usually introduced to measure the randomness and the degree of uncertainty in the system:

$$H = -\sum p_k \; log \; p_k \quad (4)$$

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 4 to calculate entropy.

C Generation Statistics

Table 4 shows generation statistics. Claim Length is the same for all settings since we sampled same 100 claims for all experiments. *Machine* tends to extract more evidences than for *Human* and it also generates longer explanations.

D Human Annotation Details

Annotation Procedure The annotation contains two tasks. The first task provides annotators with selected evidence passages asks to find masked citation sentence in the explanation. The second task requires annotators to judge the utility of the explanation in clarifying the claim (Section 3).

Annotator Recruitment The annotation task was performed on Amazon Mechanical Turk (AMT) [9]. We applied pre-screening pilot studies to find qualified annotators. We conducted individual reviews of submitted annotation results and offered feedback to annotators to address any misconceptions or confusion about the task. Annotators who performed well in the pilot study were selected to participate in final human evaluation. In order to maintain high quality throughout the annotation, we deploy quality control to identify and remove extremely poor-performing annotators. We also maintained continuous communication with annotators to address questions. In total, 68 annotators participated in the final human evaluation.

Quality Control Quality control was implemented for both pilot study and the main task. In pilot study, each Human Intelligence Task (HIT)

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

Setting	Model	Claim Length	Evidence Size	Explanation Length
Human	GPT4 GPT35 LLaMA2-70b	22.32±10.89	3.16±1.50	140.58±39.18
Machine	GPT4	22.32±10.89	5.07±2.42	171.26±51.45
	GPT35		4.68±2.58	175.89±56.99
	LLaMA2-70b		6.24±3.39	214.02±98.27

Table 4: Statistics of the data and generated explanations. The "Setting" column indicates whether evidence is human- or machine-selected. Claim Length and Explanation Length refer to token length, which is tokenized by OpenAI's tiktoken v0.7.0. Evidence Size refers to the number of selected evidence passages.

containing 3/6 control questions. 2 of them are positive questions containing exact one answer each and 1 is 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 are manually inspected to avoid any confusion. An annotator who fails on any control questions will be disqualified from participating further tasks. In total, 9 pilot studies were released to recruit qualified annotators.

During the final human evaluation, as annotators had already participated in the pilot study, a different batch of control questions were deployed to mitigate potential bias. We used ratio of failing control questions as an indicator to identify and remove extremely poor-performing workers.

Annotator Compensation Annotators were compensated $1.81 (US Dollar) per claim, and $0.38 (US Dollar) per query response pair for responses without citations. On average, annotators took approximately 10-12 minutes to complete two tasks.

Annotator Agreement During human evaluation, each claim was annotated by 5 human workers. Since common metrics like Cohen's kappa (McHugh, 2012) and Fleiss's kappa are not applicable to our multi-label annotation scenario, we used Measuring Agreement on Set-valued Items distance (MASI) (Passonneau, 2006) together with Krippendorff's alpha (Krippendorff, 2011) to compute annotator agreement. Finally, we achieved alpha value of 0.48 for GPT-4, 0.32 for GPT-35 and 0.33 for LLaMA2-70b instances.

E Prompts

Inspired by Gao et al. (2023), we developed and optimized prompt for our task. We show evidence selection prompt in Table 6 and explanation generation prompt in Table 5.

F 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.

$$\mu_i \sim \mathcal{N}(0, 1) \tag{5}$$

- Then for each annotator j, we draw their accuracy τ from a shared Gamma prior with shape parameter k and rate parameter θ [10].

$$\tau_j \sim \mathcal{G}(k, \theta) \tag{6}$$

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

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 5: Prompts for explanation generation

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

$$s_{i,j} \sim \mathcal{N}(\mu_i, \tau^{-1}) \tag{7}$$

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

$$\begin{aligned} P(s) &= \int_j P(\tau_j) \int_i P(\mu_i) P(s_{i,j}|\mu_i, \tau) d\tau d\mu \\ &= \int_j \Gamma(\tau_j|k, \theta) \int_i \mathcal{N}(\mu_i|0, 1) \\ &\qquad \mathcal{N}(s_{i,j}|\mu_j, \tau_j^{-1}) d\tau d\mu \end{aligned} \tag{8}$$

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 [11].

G Experiment Details

We ran all offline models on 4 Nvidia A100 GPUs in a data parallel fashion. Explanation generation with LLaMA2-7B takes around 2 hours and LLaMA2-70B takes around 4 hours. The cost of generating explanations was $49 (US Dollar) with GPT-4 and $10 (US Dollar) with GPT-35.

H Scientific Artifacts

We list the licenses of different artifacts used in this paper: PolitiHop (MIT) [12], Huggingface Transformers (Apache License 2.0) [13], and spaCy (MIT) [14]. Our source code and annotated data will be under MIT license.

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

[12] https://github.com/copenlu/politihop

[13] https://github.com/huggingface/transformers

[14] https://github.com/explosion/spaCy

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 6: Prompts for model evidence selection

I 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 2: 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 3: Annotation interface Page 1

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 4: Annotation interface Page 2

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.

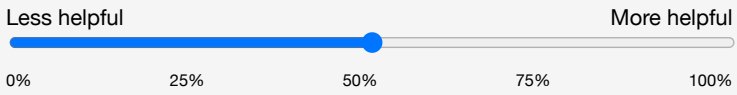
Conclusion

TRUE

Rating

Figure 5: Annotation interface Page 3

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



[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 6: Annotation interface Page 4

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.

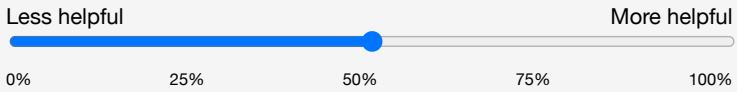
Conclusion

TRUE

Figure 7: Annotation interface Page 5

Rating

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



[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 8: Annotation interface Page 6