FACTIFY-5WQA: 5W Aspect-based Fact Verification through Question Answering

Anku Rani[1] **S.M Towhidul Islam Tonmoy**[2] **Dwip Dalal**[3]
Shreya Gautam[4] **Megha Chakraborty**[1]
Aman Chadha†[5,6] **Amit Sheth**[1] **Amitava Das**[1]
[1]University of South Carolina, USA [2]IUT, Bangladesh [3]IIT Gandhinagar, India
[4]BIT Mesra, India [5]Stanford University, USA [6]Amazon AI, USA
arani@mailbox.sc.edu amitava@mailbox.sc.edu

Abstract

Automatic fact verification has received significant attention recently. Contemporary automatic fact-checking systems focus on estimating truthfulness using numerical scores which are not human-interpretable. A human fact-checker generally follows several logical steps to verify a verisimilitude claim and conclude whether it's truthful or a mere masquerade. Popular fact-checking websites follow a common structure for fact categorization such as *half true, half false, false, pants on fire*, etc. Therefore, it is necessary to have an aspect-based (*delineating which part(s) are true and which are false*) explainable system that can assist human fact-checkers in asking relevant questions related to a fact, which can then be validated separately to reach a final verdict. In this paper, we propose a 5W framework (*who, what, when, where, and why*) for question-answer-based fact explainability. To that end, we present a semi-automatically generated dataset called FACTIFY-5WQA, which consists of 391,041 facts along with relevant 5W QAs – underscoring our major contribution to this paper. A semantic role labeling system has been utilized to locate 5Ws, which generates QA pairs for claims using a masked language model. Finally, we report a baseline QA system to automatically locate those answers from evidence documents, which can serve as a baseline for future research in the field. Lastly, we propose a robust fact verification system that takes paraphrased claims and automatically validates them. The dataset and the baseline model are available at https://github.com/ankuranii/acl-5W-QA

1 Fact checking demands aspect-based explainability

Manual fact-checking is a time-consuming task. To assess the truthfulness of a claim, a journalist would either need to search online, offline, or both, browsing through a multitude of sources while also accounting for the perceived reliability of each source. The final verdict can then be obtained via assimilation and/or comparison of the facts derived from said sources. This process can take professional fact-checkers several hours or days (Hassan et al., 2019) (Adair et al., 2017), depending on the inherent complexity of the claim.

There are several contemporary practices that journalists use for the manual verification of a claim. These methods can be categorized into four broad categories (Posetti et al., 2018):

1. **Research and fact-checking**: This involves carefully researching the claim and verifying its accuracy using reliable and credible sources such as news services, academic studies, and government data.
2. **Interviews and expert opinions**: This involves speaking with experts in the relevant field and asking for their opinions on the claim to see if it is supported by evidence and expertise.
3. **Cross-checking with multiple sources**: This involves comparing the claim with information from multiple sources to see if it is consistent or triangulates the facts obtained via multiple sources.
4. **Verifying the credibility of sources**: This involves checking the credibility of the sources used to support the claim, such as ensuring that they are reliable and unbiased.

Overall, these methods can help journalists to carefully verify claims and ensure that they are accurate and supported by evidence. However, this process is tedious and hence time-consuming. A system that can generate relevant question-answer sets by dissecting the claim into its constituent components for a given verisimilitude claim could be a great catalyst in the fact-checking process.

Research on automatic fact-checking has recently received intense attention (Yang et al.,

†Work does not relate to the position at Amazon.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10421–10440
July 9-14, 2023

Factify Question Answering at a glance					
Entailment Classes	**Textual support**	**No. of claims**	**No. of paraphrased claims**	**5WQA pairs**	**No. of evidence documents**
Support	Text are supporting each other ∼ similar news	217,856	992,503	464,766	217,635
Neutral	Text are neither supported nor refuted ∼ may have common words	79,318	365,593	194,635	45,715
Refute	Fake Claim	93,867	383,035	243,904	93,766
Total		391,041	1,741,131	903,305	357,116

Table 1: A top-level view of Factify-5WQA: (i) classes and their respective textual support specifics, (ii) Number of claims,(iii) Number of paraphrased claims, (iv) 5WQA pairs, and (v) evidence documents

5W QA based Explainability				
Claim: Moderna's lawsuits against Pfizer-BioNTech show COVID-19 vaccines were in the works before the pandemic started.				
Who claims	**What claims**	**When claims**	**Where claims**	**Why claims**
• **Q1**: ***Who lawsuits against whom?*** Ans: Moderna lawsuits against Pfizer-BioNTech	• **Q1**: ***What the lawsuit shows?*** Ans: COVID-19 vaccines were in the works before the pandemic started	• **Q1**: ***When the COVID-19 vaccines were in work?*** Ans: before pandemic.	• no claim!	• no claim!
verified true	verified false	verified false	not verifiable	not verifiable
Evidence				
• ***Moderna is suing Pfizer and BioNTech for patent infringement***, alleging the rival companies used key parts of its mRNA technology to develop their COVID-19 vaccine. Moderna's patents were filed between 2010 and 2016.	• Although the patents existed before the pandemic began, ***this does not mean Moderna or Pfizer-BioNTech were already working on the COVID-19 vaccine***. Scientists have used mRNA technology to study other viruses, such as the flu, Zika and rabies.	• Although the patents existed before the pandemic began, ***this does not mean Moderna or Pfizer-BioNTech were already working on the COVID-19 vaccine***. Scientists have used mRNA technology to study other viruses, such as the flu, Zika and rabies.	• no mention about where in any related document!	• Moderna and Pfizer-BioNTech both used messenger RNA technology, or mRNA technology, to develop their COVID-19 vaccines mention where in any related document! • This technology dates back to the 1990s, but the first time mRNA vaccines were widely disseminated was to combat the spread of COVID-19.

Table 2: An illustration of 5W QA-based explainable fact verification system. This example is an illustration of the false claim. A typical semantic role labeling (SRL) system processes a sentence and identifies verb-specific semantic roles. Therefore, for the specified example, we have one sentence that has two main verbs *were, and started*. For each verb, 5W QA pair will automatically be generated ($2 \times 5 = 10$) 10 sets of QA pairs in total for this example. Further, all those 10 5W aspects will be fact-checked using evidence. If in case of some aspects ended having *neutral* entailment verdict, possible relevant documents with associated URLs will be listed for the end user to further read and assess. This will aid human fact-checkers.

2022a), (Park et al., 2021), (Atanasova et al., 2019), (Guo et al., 2022), (Trokhymovych and Saez-Trumper, 2021). Several datasets to evaluate automatic fact verification such as FEVER (Thorne et al., 2018a), LIAR (Wang, 2017), PolitiFact (Garg and Sharma, 2020), FavIQ (Kwiatkowski et al., 2019), Hover (Jiang et al., 2020), X-Fact (Gupta and Srikumar, 2021), CREAK (Onoe et al., 2021), FEVEROUS (Aly et al., 2021) are also available.

Contemporary automatic fact-checking systems focus on estimating truthfulness using numerical scores which are not human-interpretable (Nakov et al., 2021; Guo et al., 2021). Others extract explicit mentions of the candidate's facts in the text as evidence for the candidate's facts, which can be hard to spot directly. Moreover, in the case of false information, it is commonplace that the whole claim isn't false, but some parts of it are, while others could still be true. A claim is either opinion-based, or knowledge-based (Kumar and Shah, 2018). For the same reason, the popular website Politifact based on the work by (Garg and Sharma, 2020) categorized the fact-checking verdict in the form of half-true, half-false, etc.

We propose 5W (Who, What, When, Where, and Why) aspect-based question-answer pairwise explainability. Including these 5W elements within a statement can provide crucial information regarding the entities and events being discussed, thus facilitating a better understanding of the text. For instance, in the statement "*Moderna's lawsuits against Pfizer-BioNTech show COVID-19 vaccines were in the works before the pandemic started.*" The use of *who* highlights the individuals or entities involved in the action of filing lawsuits, *what* pertains to the content of the lawsuit, specifically the revelation that COVID-19 vaccines were in the works, *when* refers to the timing of this revelation, i.e., before the pandemic. Overall, the incorporation of "who," "what," "when," "where," and "why" in a text can provide crucial context and aid in making the text more clear and comprehensible.

Automatic question and answering (Q&A) systems can provide valuable support for claims by providing evidence and supporting information. They can also help to identify potential flaws or weaknesses in a claim, allowing for further analysis and discussion. They can also help to identify potential flaws or weaknesses in a claim, allowing for further analysis and discussion.

Only two recent works (Yang et al., 2022b; Kwiatkowski et al., 2019) propose question answering as a proxy to fact verification explanation, breaking down automated fact-checking into several steps and providing a more detailed analysis of the decision-making processes. Question-answering-based fact explainability is indeed a very promising direction. However, open-ended QA for a fact can be hard to summarize. Therefore, we refine the QA-based explanation using the 5W framework (*who, what, when, where, and why*). Journalists follow an established practice for fact-checking, verifying the so-called 5Ws (Mott, 1942), (Stofer et al., 2009), (Silverman, 2020), (Su et al., 2019), (Smarts, 2017). This directs verification search and, moreover, identifies missing content in the claim that bears on its validity. One consequence of journalistic practice is that claim rejection is not a matter of degree (*as conveyed by popular representations such as a number of Pinocchios or crows, or true, false, half true, half false, pants on fire*), but the rather specific, substantive explanation that recipients can themselves evaluate (Dobbs, 2012).

2 Data sources and compilation

Data collection is done by sorting 121 publicly available prevalent fact verification data sets based on modalities (111), languages (83), and tasks (51).

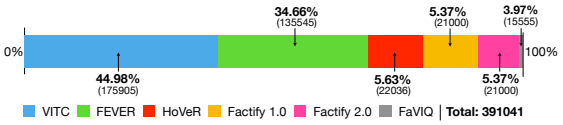


Figure 2: Distribution of the FACTIFY 5WQA fact verification dataset.

By filtering 121 publicly available data sets for fact verification, we found ten of them to be suitable for the text-based fact verification task. We only considered the claims present in textual format in English-language because of which DanFEVER (Nørregaard and Derczynski, 2021) and X-Fact (Gupta and Srikumar, 2021) were also excluded because they are either Danish or multilingual. We discovered that "Evidence-based Factual Error Correction" and FEVEROUS (Aly et al., 2021) were subsets of the FEVER dataset, so we decided to use FEVER (Thorne et al., 2018b), HoVer (Jiang et al., 2020), VITC (Schuster et al., 2021), FaVIQ (Park et al., 2021), Factify 1.0 (Patwa et al., 2022) and Factify 2.0 (Mishra et al., 2022) for our analysis. We verified that the claims in these datasets were unique but found that 64 claims from VITC (Schuster et al., 2021) overlapped with those in FEVER (Thorne et al., 2018b) which is later considered once giving a total count of $391,041$ datapoints and the distribution is represented in the figure 2.

We only used a specific number of claims from each of the six datasets after manually inspecting the quality aspects - length of the claim and evidence, grammatical correctness, etc. For the FEVER and VITC datasets, only the claims belonging to the train split were used for making the dataset. For Factify 1.0 and Factify 2.0, the multimodal part of the dataset was discarded and only the text-based part was used. FaVIQ has two sets: the *A set* and the *R set*. *A set* consists of ambiguous questions and their disambiguation. *R set* is made by using unambiguous question-answer pairs. As discussed in earlier paragraphs, *A set* is a more challenging set; hence we took the *A set* of FaVIQ for making our dataset. In the case of the HoVer dataset, 22036 claims were used in making our dataset.

We propose an amalgamated data set with the total number of unique claims as $391,041$. Around ($\sim$ 85%) of them are from VITC, FEVER, and

(a) HoVer (b) FaVIQ (c) VITC

(d) FEVER (e) Factify 1.0 (f) Factify 2.0

Figure 1: Word cloud offers a glance view of topic distributions over the chosen datasets: (i) VITC (Schuster et al., 2021), (ii) FEVER (Thorne et al., 2018b), (iii) Factify 1.0 (Patwa et al., 2022), (iv) Factify 2.0 (Mishra et al., 2022), (v) HoVer (Jiang et al., 2020), and (vi) FaVIQ (Park et al., 2021). Darker color shades in the cloud represent a higher frequency of the particular word in the dataset.

HoVer, and ($\sim 15\%$) of it is from Factify 1.0, Factify 2.0 and FaVIQ as evident from Figure 2. Figure 1 offers a snapshot of topics in these datasets through a word cloud.

3 Paraphrasing textual claims

The motivation behind paraphrasing textual claims is as follows. A textual given claim may appear in various different textual forms in real life, owing to variations in the writing styles of different news publishing houses. Incorporating such variations is essential to developing a strong benchmark to ensure a holistic evaluation (see examples in Figure 3).Manual generation of possible paraphrases is undoubtedly ideal, but that process is time-consuming and labor-intensive. On the other hand, automatic paraphrasing has received significant attention in recent times (Niu et al., 2020) (Nicula et al., 2021)(Witteveen and Andrews, 2019)(Nighojkar and Licato, 2021). For a given claim, we generate multiple paraphrases using various SoTA models. In the process of choosing the appropriate paraphrase model based on a list of available models, the primary question we asked is how to make sure the generated paraphrases are rich in diversity while still being linguistically correct. We delineate the process followed to achieve this as follows. Let's say we have a claim c. We generate n paraphrases using a paraphrasing model. This yields a set of $p_1^c, \ldots, p_n^c$. Next, we make pair-wise comparisons of these paraphrases with c, resulting in $c - p_1^c, \ldots$, and $c - p_n^c$. At this step, we identify the

Moderna's lawsuits against Pfizer-BioNTech show COVID-19 vaccines were in the works before the pandemic started.

Prphr 1: Moderna's legal action against Pfizer-BioNTech demonstrates that work was being done on COVID-19 vaccines prior to the outbreak of the pandemic.

Prphr 2: Moderna's legal action against Pfizer-BioNTech implies that work on COVID-19 vaccines had begun prior to the beginning of the pandemic.

Prphr 3: Moderna's court cases against Pfizer-BioNTech indicate that COVID-19 vaccines had been in development before the pandemic began.

Prphr 4: Moderna's prosecution against Pfizer-BioNTech demonstrates that COVID-19 vaccines had been in advancement prior to the pandemic commencing.

Prphr 5: It is revealed by Moderna's legal actions addressed to Pfizer-BioNTech that work on COVID-19 vaccines was being done before the pandemic began.

Figure 3: Claims and paraphrases obtained using `text-davinci-003` (Brown et al., 2020)

examples which are entailed, and only those are chosen. For the entailment task, we have utilized RoBERTa Large (Liu et al., 2019) – a SoTA model trained on the SNLI task (Bowman et al., 2015).

However, there are many other secondary factors, for e.g., a model may only be able to generate a limited number of paraphrase variations compared to others, but others can be more correct and/or consistent. As such, we considered three major dimensions in our evaluation: *(i) a number of considerable paraphrase generations, (ii) correctness in those generations, and (iii) linguistic diversity in those generations*. We conducted experiments with three available models: (a) Pegasus (Zhang et al., 2020), (b) T5 (T5-Large) (Raffel et al., 2020), and (c) GPT-3 (`text-davinci-003` variant) (Brown et al., 2020). Based on empirical observations, we concluded that GPT-3 outperformed all the other models. To offer transparency around our experiment process, we detail the aforementioned evaluation dimensions as follows.

Model	Coverage	Correctness	Diversity
Pegasus	32.46	94.38%	3.76
T5	30.26	83.84%	3.17
GPT-3	35.51	88.16%	7.72

Table 3: Experimental results of automatic paraphrasing models based on three factors: *(i) coverage, (ii) correctness and (iii) diversity*; GPT-3 (`text-davinci-003`) can be seen as the most performant.

Coverage - a number of considerable paraphrase generations: We intend to generate up to 5 paraphrases per given claim. Given all the generated claims, we perform a minimum edit distance (MED) (Wagner and Fischer, 1974) - units are words instead of alphabets). If MED is greater than ± 2 for any given paraphrase candidate (for e.g., $c - p_1^c$) with the claim, then we further consider that paraphrase, otherwise discarded. We evaluated all three models based on this setup that what model is generating the maximum number of considerable paraphrases.

Correctness - correctness in those generations: After the first level of filtration we have performed pairwise entailment and kept only those paraphrase candidates, are marked as entailed by the (Liu et al., 2019) (Roberta Large), SoTA trained on SNLI (Bowman et al., 2015).

Diversity - linguistic diversity in those generations: We were interested in choosing that model can produce linguistically more diverse paraphrases. Therefore we are interested in the dissimilarities check between generated paraphrase claims. For e.g., $c - p_n^c$, $p_1^c - p_n^c$, $p_2^c - p_n^c$, ..., $p_{n-1}^c - p_n^c$ and repeat this process for all the other paraphrases and average out the dissimilarity score. There is no such metric to measure dissimilarity, therefore we use the inverse of the BLEU score (Papineni et al., 2002). This gives us an understanding of how linguistic diversity is produced by a given model. Based on these experiments, we found that `text-davinci-003` performed the best. The results of the experiment are reported in the following table. Furthermore, we were more interested to choose a model that can maximize the linguistic variations, and `text-davinci-003` performs on this parameter of choice as well. A plot on diversity vs. all the chosen models is reported in Figure 4.

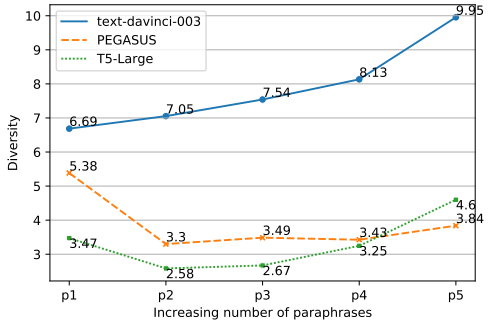


Figure 4: A higher diversity score depicts an increase in the number of generated paraphrases and linguistic variations in those generated paraphrases.

4 5W Semantic Role Labelling

Claim.	Moderna's lawsuits against Pfizer-BioNTech show COVID-19 vaccines were in the works before the pandemic started.
Verbs.	[were, started]
Who.	[(ARG0 : None, None)]
What.	[(ARG1 : COVID-19 vaccines, the pandemic)]
Why.	[(ARGM-TMP : None, None)]
When.	[(ARGM-LOC : before the pandemic started, None)]
Why.	[(ARGM-TMP : None, None)]
Where.	[(ARGM-CAU : None, None)]

Figure 5: Examples of the 5W semantic role labels.

Identification of the functional semantic roles played by various words or phrases in a given sentence is known as semantic role labelling (SRL). SRL is a well-explored area within the NLP community. There are quite a few off-the-shelf tools available: (i) Stanford SRL (Manning et al., 2014), (ii) AllenNLP (AllenNLP, 2020), etc. A typical SRL system first identifies verbs in a given sentence and then marks all the related words/phrases haven relational projection with the verb and assigns appropriate roles. Thematic roles are gener-

ally marked by standard roles defined by the Proposition Bank (generally referred to as PropBank) (Palmer et al., 2005), such as: *Arg0, Arg1, Arg2*, and so on. We propose a mapping mechanism to map these PropBank arguments to 5W semantic roles. (look at the conversion table 4).

Semantic role labelling (SRL) is a natural language processing technique that involves identifying the functions of different words or phrases in a sentence. This helps to determine the meaning of the sentence by revealing the relationships between the entities in the sentence. For example, in the sentence "*Moderna's lawsuits against Pfizer-BioNTech show COVID-19 vaccines were in the works before the pandemic started,*" *Moderna* would be labeled as the *agent* and *Pfizer-BioNTech* would be labelled as the *patient*.

PropBank Role	Who	What	When	Where	Why	How
ARG0	**84.48**	0.00	3.33	0.00	0.00	0.00
ARG1	10.34	**53.85**	0.00	0.00	0.00	0.00
ARG2	0.00	9.89	0.00	0.00	0.00	0.00
ARG3	0.00	0.00	0.00	22.86	0.00	0.00
ARG4	0.00	3.29	0.00	34.29	0.00	0.00
ARGM-TMP	0.00	1.09	**60.00**	0.00	0.00	0.00
ARGM-LOC	0.00	1.09	10.00	**25.71**	0.00	0.00
ARGM-CAU	0.00	0.00	0.00	0.00	**100.00**	0.00
ARGM-ADV	0.00	4.39	20.00	0.00	0.00	0.06
ARGM-MNR	0.00	3.85	0.00	8.57	0.00	**90.91**
ARGM-MOD	0.00	4.39	0.00	0.00	0.00	0.00
ARGM-DIR	0.00	0.01	0.00	5.71	0.00	3.03
ARGM-DIS	0.00	1.65	0.00	0.00	0.00	0.00
ARGM-NEG	0.00	1.09	0.00	0.00	0.00	0.00

Table 4: A mapping table from PropBank(Palmer et al., 2005) (*Arg0, Arg1, ...*) to 5W (*who, what, when, where, and why*).

The five "W"s (what, when, where, why, who) are often used to refer to the key questions that need to be answered in order to fully understand a sentence or piece of text. SRL can be seen as a way of providing answers to these questions by identifying the various roles that words and phrases play within a sentence. For example, a semantic role labeler might identify the subject of a sentence (who or what the sentence is about), the object (who or what is being acted upon), and the verb (the action being performed). In this way, semantic role labeling can be seen as a way of providing the necessary context for answering the five "W"s, and can be an important tool in natural language processing and understanding.

In this study, we use the mapping displayed in table 4 and replace the roles that are assigned with respect to each verb as an output from SRL with 5W. According to table 4, it is evident that each of the 5Ws can be mapped to semantic roles. The highest percentage of mapping is taken into consideration and concluded in table 4.

After the mapping is done, a detailed analysis for the presence of each of the 5W is conducted which is summarized in figure 6.

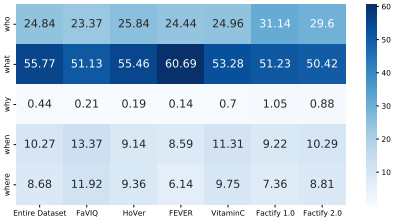


Figure 6: Percentage of W's present across the dataset.

In this study, experimentation for finding semantic roles was conducted using **AllenNLP SRL** demo (AllenNLP, 2020). Developed by (Shi and Lin, 2019), it is a BERT (Devlin et al., 2018) based model with some modifications that introduce a linear classification layer with no additional parameters, and it is currently the best single model for English PropBank SRL on newswire sentences with a test F1 of 86.49 on the Ontonotes 5.0 dataset (Palmer et al., 2005). Newswire instances correlate with the fact verification dataset as true news is also a fact.

As indicated in figure 5, the pipeline for generating 5W aspect-based semantic role labeling is to pass it through an SRL model and map it with 5W. An example of a claim as per the output using AllenNLP's SRL model is in figure 5.

4.1 Human Evaluation of the 5W SRL

In this work evaluation for the 5W Aspect, based on semantic role labeling is conducted using *mapping accuracy*: This involves accuracy on SRL output mapped with 5Ws.

For the purpose of finding how good the mapping of 5W is with semantic roles and generation of semantic roles, human annotation of 3000 data points was conducted. 500 random data points each from FEVER , FavIQ, HoVer, VITC, Factify 1.0 and Factify 2.0 were annotated and the results are described in table 6.

	FaVIQ	FEVER	HoVer	VitaminC	Factify 1.0	Factify 2.0
Who	89%	85%	90%	87%	86%	82%
What	85%	56%	68%	78%	81%	93%
When	86%	90%	95%	98%	83%	75%
Where	93%	100%	90%	97%	93%	86%
Why	0%	-	100%	92%	87%	93%

Table 6: Human evaluation of 5W SRL; % represents human agreement on 5W mapping with SRL.

5 5W aspect-based QA pair generation

A false claim is very likely to have some truth in it, some correct information. In fact, most fake news

	ProphetNet								BART							
	Claim				+Paraphrase				Claim				+Paraphrase			
	BLEU	ROUGE-L	Recall	F1	BLEU	ROUGE-L	Recall	F1	BLEU	ROUGE-L	Recall	F1	BLEU	ROUGE-L	Recall	F1
T5-3b	**29.22**	**48.13**	**35.66**	**38.03**	**28.13**	**46.18**	**34.15**	**36.62**	21.78	34.53	28.03	28.07	20.93	33.57	27.65	27.24
T5-Large	28.81	48.02	35.26	37.81	21.46	46.45	27.19	36.76	21.46	34.90	27.41	27.99	20.88	33.69	20.88	27.31
BERT large	28.65	46.25	34.55	36.72	27.27	44.10	32.95	35	20.66	33.19	25.51	26.44	19.74	32.34	25.14	25.71

Table 5: Selecting the best combination - 5W QAG vs. 5W QA validation.

articles are challenging to detect precisely because they are mostly based on correct information, deviating from the facts only in a few aspects. That is, the misinformation in the claim comes from a very specific inaccurate statement. So, given our textual claim, we generate 5W question-answer pairs by doing semantic role labeling on the given claim. The task is now based on the generated QA pairs, a fact-checking system can extract evidence sentences from existing authentic resources to verify or refute the claim based on each question- *Who, What, When, Where, and Why* (Wikipedia, 2023). Please see examples in Figure 7.

Claim.	Moderna's lawsuits against Pfizer-BioNTech show COVID-19 vaccines were in the works before the pandemic started
QA Pair1.	[(Who lawsuits against whom?, Moderna lawsuits against Pfizer-BioNTech)]
QA Pair2.	[(What the law- suit shows?, COVID-19 vaccines were in the works before the pandemic started.)]
QA Pair3.	[(When the COVID-19 vaccines were in work?, before pandemic.)]

Figure 7: Examples of QA pairs generated from a claim by the QG system.

Our method of using 5W SRL to generate QA pairs and then verify each aspect separately allows us to detect '*exactly where the lie lies*'. This, in turn, provides an explanation of why a particular claim is refutable since we can identify exactly which part of the claim is false.

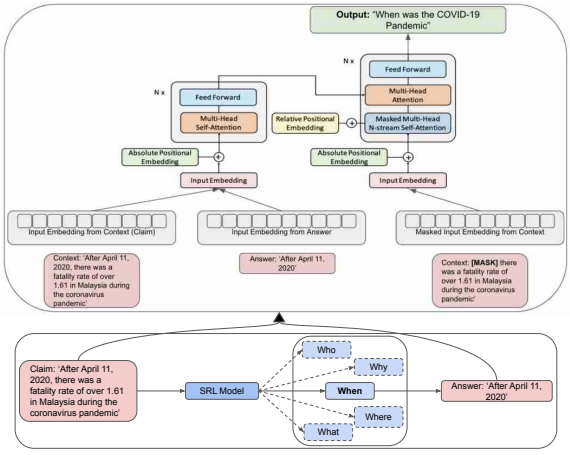


Figure 8: Illustration of 5W QA Generation Pipeline using ProphetNet.

The process of fact verification is inherently intricate, with several questions representing the components within the underlying claim that need answers to reach a verdict on the veracity of the claim. Referring to the example in figure 7, such questions may include: *(a) Who lawsuit against whom? (b) Vaccine were in use when?* what can go wrong if this claim is false? Manual fact-checking can be labor-intensive, consuming several hours or days (Hassan et al., 2015; Adair et al., 2017).

For the 5W question generation task we have experimented with two models: (i) BART (Lewis et al., 2019), and (ii) ProphetNet (Qi et al., 2020), and found ProphetNet outperforms the former.

ProphetNet (Qi et al., 2020), a generative model that uses multi-lingual pre-training with masked span generation. It is optimized through *n-step* ahead prediction, which predicts the next *n* tokens based on previous context tokens at each time step, encouraging the model to explicitly plan for future tokens. In this work, we employed the context-based question generation approach to generate relevant and specific questions for the task of fact verification. This approach utilizes the claim information to ensure that the generated questions are appropriate for fact-checking.

5.1 Human evaluation of QA generation

		FaVIQ	FEVER	HoVer	VitaminC	Factify 1.0	Factify 2.0
Who	Question is well-formed	86%	77%	84%	79%	80%	82%
	Question is correct	90%	82%	86%	83%	87%	89%
	Answer is correct	89%	85%	90%	87%	86%	82%
What	Question is well-formed	71%	53%	68%	79%	77%	72%
	Question is correct	77%	69%	70%	81%	80%	76%
	Answer is correct	85%	56%	68%	78%	81%	93%
When	Question is well-formed	88%	77%	86%	78%	81%	78%
	Question is correct	90%	86%	88%	94%	92%	89%
	Answer is correct	86%	90%	95%	98%	83%	75%
Where	Question is well-formed	90%	95%	68%	87%	91%	88%
	Question is correct	85%	95%	78%	92%	92%	83%
	Answer is correct	93%	97%	90%	97%	93%	86%
Why	Question is well-formed	0%	-	100%	92%	92%	90%
	Question is correct	0%	-	100%	95%	95%	94%
	Answer is correct	0%	-	100%	96%	87%	93%

Table 7: Human evaluation of QA generation. % represents human agreement on how well the question is formed, and whether the question and answer are correct.

For the evaluation purpose, a random sample of 3000 data points was selected for annotation. The questions generated using the Prophetnet model were utilized for this purpose. The annotators were instructed to evaluate the question-answer pairs in three dimensions: the question is well formed,

which means it is syntactically correct, the question is correct which means it is semantically correct with respect to the given claim, and extracted answer from the model is correct. The evaluation results for the datasets are presented in the following analysis.

6 The 5W QA validation system

Finally, we propose a QA validation system, where the generated questions from the QG system and the evidence are passed through SoTA Question answering models (T5:3B) (Raffel et al., 2020), T5:Large (Raffel et al., 2020), Bert: Large (Devlin et al., 2018)) demonstrated in figure 9. This helps to find out whether the evidence supports or refutes the claim or if the system misses out on enough information to make a conclusion.

Claim:	Moderna's lawsuits against Pfizer-BioNTech show COVID-19 vaccines were in the works before the pandemic started.
Evidence:	Moderna is suing Pfizer and BioNTech for patent infringement, alleging the rival companies used key parts of its mRNA technology to develop their COVID-19 vaccine. Although the patents existed before the pandemic began, this does not mean Moderna or Pfizer-BioNTech were already working on the COVID-19 vaccine.
Model Generated Question1:	Who lawsuits against whom?,
Gold Answer:	Moderna lawsuits against Pfizer-BioNTech.
Answer Generated from System:	Moderna
Model Generated Question2:	What the lawsuit shows?,
Gold Answer:	COVID-19 vaccines were in the works before the pandemic started.
Answer Generated from System:	Patent infringement
Model Generated Question3:	When the COVID-19 vaccines were in work?,
Gold Answer:	Before pandemic.
Answer Generated from System:	Before the pandemic began.

Figure 9: Examples of QA pairs generated from evidence by the QA system.

An example of two of the claims that generate answers based on the evidence is represented in figure 9. In this figure, the question is generated using prophetnet, and the answer is generated using the T5-3B model from the evidence of the claims. as described in figure 10.

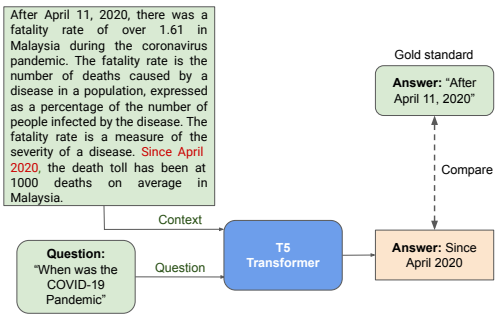


Figure 10: T5-based question answering framework.

To design the 5W QA validation system, we utilized the claims, evidence documents, and 5W questions generated by the question generation system as input. The answer generated by the 5W QG model is treated as the gold standard for comparison between claim and evidence. We experimented with three models, T5-3B (Raffel et al., 2020), T5-Large (Raffel et al., 2020), and Bert-Large (Devlin et al., 2018). The T5 is an encoder-decoder-based language model that treats this task as text-to-text conversion, with multiple input sequences and produces an output as text. The model is pre-trained using the C4 corpus (Raffel et al., 2020) and fine-tuned on a variety of tasks. T5-Large employs the same encoder-decoder architecture as T5-3B (Raffel et al., 2020), but with a reduced number of parameters. The third model that we experimented with is the Bert-Large (Devlin et al., 2018) model, which utilizes masked language models for pre-training, enabling it to handle various downstream tasks.

7 Selecting the best combination - 5W QAG vs. 5W QA validation

We have utilized off-the-self models both for 5W question-answer generation and 5W question-answer validation. Given that the datasets used for training the models bear an obvious discrepancy in terms of the distribution characteristics compared to our data (world news) which would probably lead to a generalization gap, it was essential to experimentally judge which system offered the best performance for our use-case. Instead of choosing the best system for generation vs. validation, we opted for pair-wise validation to ensure we chose the best combination. Table 5 details our evaluation results – the rows denote the QA models while the columns denote QAG models. From the results in the table, we can see that the best combination in terms of a QAG and QA validation model was identified as T5-3b and ProphetNet, respectively.

8 Conclusion and future avenues

It has been realized by the community that due to the given complexity of fact-checking it possibly can not be automated completely. Human-in-loop is the solution for the same. Proposed 5W QA-based fact verification can be the best aid for human fact-checkers. To the best of our knowledge, we are the first to introduce 5W QA-based fact verification and additionally proposed relevant techniques to automatically generate QA using the automatic method, which can be readily used for any incoming claim on the spot. Furthermore, the QA validation section can aid to provide evidence

support. Paraphrasing claims provide a holistic approach to fact-checking. Generated datasets and resources will be made public for research purposes containing 3.91 million claims.

9 Discussion and limitations

In this section, we self-criticize a few aspects that could be improved and also detail how we plan (tentatively) to plan to improve upon those specific aspects -

9.1 Paraphrasing claims

Manual generation of possible paraphrases is undoubtedly ideal but is time-consuming and labor-intensive. Automatic paraphrasing is a good way to scale quickly, but there could be more complex variations of meaning paraphrases hard to generate automatically. For example - "*It's all about business - a patent infringement case against Pfizer by a rival corporate reveals they knew about COVID in one way!*" and "*Oh my god COVID is not enough now we have to deal with HIV blood in the name of charity!*".

An ideal for this shortcoming would be to manually generate a few thousand paraphrase samples and then fine-tune language models. On the other hand, a new paradigm in-context Learning is gaining momentum (Xun et al., 2017). In-context learning has been magical in adapting a language model to new tasks through just a few demonstration examples without doing gradient descent. There are quite a few recent studies that demonstrate new abilities of language models that learn from a handful of examples in the context (in-context learning - ICL for short). Many studies have shown that LLMs can perform a series of complex tasks with ICL, such as solving mathematical reasoning problems (Wei et al., 2022). These strong abilities have been widely verified as emerging abilities for large language models (Wei et al., 2022). From prompt engineering to chain of thoughts, we are excited to do more experiments with the new paradigm of in-context learning for automatically paraphrasing claims.

9.2 5W SRL

Semantic role labeling is a well-studied subdiscipline, and the mapping mechanism we proposed works well in most cases except in elliptic situations like anaphora and cataphora. In the future, we would like to explore how an anaphora and coreference resolution (Joshi et al., 2019) can aid an improvement.

9.3 5W QA pair generation

5W semantic role-based question generation is one of the major contributions of this paper. While automatic generation aided in scaling up the QA pair generation, it also comes with limitations of generating more complex questions covering multiple Ws and *how* kinds of questions; for example, "*How Moderna is going to get benefited if this Pfizer COVID news turns out to be a rumor?*". For the betterment of FACTIFY benchmark, we would like to generate few thousand manually generated abstract QA pairs. Then will proceed towards in-context Learning (Xun et al., 2017).

Abstractive question-answering has received momentum (Zhao et al., 2022), (Pal et al., 2022) recently. We want to explore how we can generate more abstract QA pairs for the multimodal fact-verification task.

9.4 QA system for the 5W question

Generated performance measures attest the proposed QA model needs a lot more improvement. This is due to the complexity of the problem and we believe that will attract future researchers to try this benchmark and conduct research on multimodal fact verification.

It has been realized by the community that relevant document retrieval is the major bottleneck for fact verification. Recent work introduced a fresh perspective to the problem - named Hypothetical Document Embeddings (HyDE) (Gao et al., 2022) and applied a clever trick even if the wrong answer is more semantically similar to the right answer than the question. This could be an interesting direction to explore and examine how that could aid in retrieving relevant documents and answers.

References

Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward "the holy grail": The continued quest to automate fact-checking. In *Computation+ Journalism Symposium,(September)*.

AllenNLP. 2020. Allennlp semantic role labeling. https://demo.allennlp.org/semantic-role-labeling. [Online; accessed 2023-01-02].

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Fever-

ous: Fact extraction and verification over unstructured and structured information.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Michael Dobbs. 2012. The rise of political fact-checking, how reagan inspired a journalistic movement. *New America Foundation*, pages 4–5.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Sonal Garg and Dilip Kumar Sharma. 2020. New politifact: A dataset for counterfeit news. In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 17–22.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Tarek A Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2019. Firm-Level Political Risk: Measurement and Effects*. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019. BERT for coreference resolution: Baselines and analysis. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal fact verification dataset. In *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*.

Frank Luther Mott. 1942. Trends in newspaper content. *The Annals of the American Academy of Political and Social Science*, 219:60–65.

Preslav Nakov, David P. A. 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. *CoRR*, abs/2103.07769.

Bogdan Nicula, Mihai Dascalu, Natalie Newton, Ellen Orcutt, and Danielle S McNamara. 2021. Automated paraphrase quality assessment using recurrent neural networks and language models. In *International Conference on Intelligent Tutoring Systems*, pages 333–340. Springer.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint arXiv:2106.07691*.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2020. Unsupervised paraphrasing with pretrained language models. *arXiv preprint arXiv:2010.12885*.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge.

Vaishali Pal, Evangelos Kanoulas, and Maarten Rijke. 2022. Parameter-efficient abstractive question answering over tables or text. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 41–53, Dublin, Ireland. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Faviq: Fact verification from information-seeking questions. *arXiv preprint arXiv:2107.02153*.

Parth Patwa, Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Benchmarking multi-modal entailment for fact verification. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Julie Posetti, Cherilyn Ireton, Claire Wardle, Hossein Derakhshan, Alice Matthews, Magda Abu-Fadil, Tom Trewinnard, Fergus Bell, and Alexios Mantzarlis. 2018. Unesco. https://unesdoc.unesco.org/ark:/48223/pf0000265552. [Online; accessed 2023-01-02].

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

Craig Silverman. 2020. Verification handbook: Homepage.

Media Smarts. 2017. How to recognize false content online - the new 5 ws.

Kathryn T Stofer, James R Schaffer, and Brian A Rosenthal. 2009. *Sports journalism: An introduction to reporting and writing*. Rowman & Littlefield Publishers.

Jing Su, Xiguang Li, and Lianfeng Wang. 2019. The study of a journalism which is almost 99% fake. *Lingue Culture Mediazioni-Languages Cultures Mediation (LCM Journal)*, 5(2):115–137.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018b. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Mykola Trokhymovych and Diego Saez-Trumper. 2021. Wikicheck: An end-to-end open source automatic fact-checking api based on wikipedia. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4155–4164.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

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

Wikipedia. 2023. Five ws.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*.

Guangxu Xun, Xiaowei Jia, Vishrawas Gopalakrishnan, and Aidong Zhang. 2017. A survey on context learning. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):38–56.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022a. Explainable fact-checking through question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8952–8956. IEEE.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022b. Explainable fact-checking through question answering. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8952–8956.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Wenting Zhao, Konstantine Arkoudas, Weiqi Sun, and Claire Cardie. 2022. Compositional task-oriented parsing as abstractive question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4418–4427, Seattle, United States. Association for Computational Linguistics.

10 FAQ

1. 5W SRL is understandable, but how is the quality of the 5W QA pair generation using a language model?

   Ans. - We have evaluated our QA generation against the SoTA model for QA Tasks - T5. Please refer to the section 7, table 5 for a detailed description of the process and evaluation. Moreover, please see the discussion in the limitation section 9.3.

2. How were models shortlisted for Question generation?

   Ans. - We have shortlisted the current SOTA models on question generation-specific tasks. Due to our resource limitation, we have gone for those models that are open-sourced, are not resource heavy, and produce great results without fine-tuning them.

3. How were models shortlisted for the question-answering system?

   Ans. - Selected the current SOTA models that have lower inference time but produce great results on text generation tasks.

4. Why was absolute value 2 chosen as a filter for minimum edit distance?

   Ans. - Edit distance is a measure of the similarity between two pieces of text, and a higher value generally indicates more diversity. A higher minimum edit distance between the input and generated text indicates that the generated text is more unique and less likely to be a simple copy or repetition of the input. Therefore, it is commonly held that a minimum edit distance of greater than 2 is a desirable characteristic in natural language generation systems.

5. How was the prompt-based paraphrasing done using the `text-davinci-003` model?

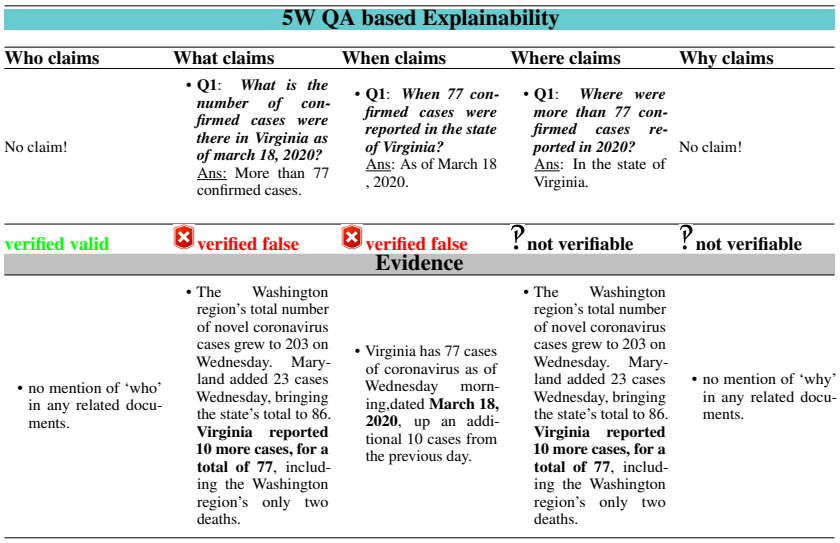
   Ans. - As `text-davinci-003` is a prompt-based model and so we had to create a prompt that would instruct `text-davinci-003` to generate five paraphrases for the given input claims. Careful consideration was given to ensure that the prompt would generate output with a specific syntax, as this was necessary for the efficient application of the model to a large number of claims. Through experimentation with multiple different prompts, we came to the conclusion that the following prompt works best:
   "Generate five different paraphrases of the following text and then place all these five paraphrases in one list of python format. Do not write anything other than just the list "
   We also developed a post-processing pipeline to ensure that if there is a slight variation in the syntax of paraphrases generated, then we can easily extract those paraphrases from the output of `text-davinci-003`.

6. How was the diversity vs. the number of paraphrases graph plotted?

   Ans. - After the two layers of filtration, i.e., filtering it by coverage and correctness, the obtained paraphrases are then used to calculate the diversity score as described in section 3. Let d_i represent the diversity score of the i^{th} paraphrase generated. So in order to get the general diversity score for the i^{th} paraphrase, we computed the average d_i score of all i^{th} paraphrases generated.

Appendix

This section provides additional examples to assist in the understanding and interpretation of the research work presented.

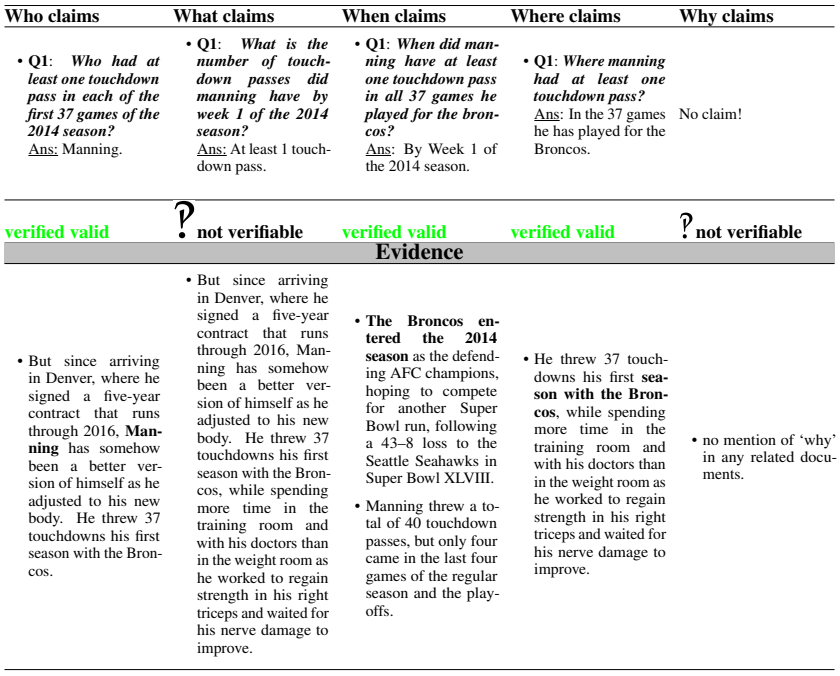


5W QA based Explainability				
Who claims	**What claims**	**When claims**	**Where claims**	**Why claims**
No claim!	• **Q1**: ***What is the number of confirmed cases were there in Virginia as of march 18, 2020?*** Ans: More than 77 confirmed cases.	• **Q1**: ***When 77 confirmed cases were reported in the state of Virginia?*** Ans: As of March 18 , 2020.	• **Q1**: ***Where were more than 77 confirmed cases reported in 2020?*** Ans: In the state of Virginia.	No claim!
verified valid	verified false	verified false	not verifiable	not verifiable
Evidence				
• no mention of 'who' in any related documents.	• The Washington region's total number of novel coronavirus cases grew to 203 on Wednesday. Maryland added 23 cases Wednesday, bringing the state's total to 86. **Virginia reported 10 more cases, for a total of 77**, including the Washington region's only two deaths.	• Virginia has 77 cases of coronavirus as of Wednesday morning,dated **March 18, 2020**, up an additional 10 cases from the previous day.	• The Washington region's total number of novel coronavirus cases grew to 203 on Wednesday. Maryland added 23 cases Wednesday, bringing the state's total to 86. **Virginia reported 10 more cases, for a total of 77**, including the Washington region's only two deaths.	• no mention of 'why' in any related documents.

As of March 18 , 2020 , there were more than 77 confirmed cases reported in the state of Virginia.
Prphr 1: According to records updated on the 18th of March 2020, the state of Virginia has more than 77 COVID-19 cases.
Prphr 2: Based on the data of March 18th, 2020, there are over 77 reported cases of coronavirus in Virginia.
Prphr 3: By March 18 2020, Virginia has a reported number of more than 77 certified cases of the coronavirus.
Prphr 4: As of the 18th of March 2020, there was evidence of 77 positive coronavirus cases in Virginia.
Prphr 5: As of March 18th 2020, 77 documented incidences of coronavirus had been raised in Virginia.

Figure 12: Claims paraphrased using `text-davinci-003`

Figure 11: Claim: As of March 18 , 2020 , there were more than 77 confirmed cases reported in the state of Virginia.



Who claims	**What claims**	**When claims**	**Where claims**	**Why claims**
• **Q1**: ***Who had at least one touchdown pass in each of the first 37 games of the 2014 season?*** Ans: Manning.	• **Q1**: ***What is the number of touchdown passes did manning have by week 1 of the 2014 season?*** Ans: At least 1 touchdown pass.	• **Q1**: ***When did manning have at least one touchdown pass in all 37 games he played for the broncos?*** Ans: By Week 1 of the 2014 season.	• **Q1**: ***Where manning had at least one touchdown pass?*** Ans: In the 37 games he has played for the Broncos.	No claim!
verified valid	not verifiable	verified valid	verified valid	not verifiable
Evidence				
• But since arriving in Denver, where he signed a five-year contract that runs through 2016, **Manning** has somehow been a better version of himself as he adjusted to his new body. He threw 37 touchdowns his first season with the Broncos.	• But since arriving in Denver, where he signed a five-year contract that runs through 2016, Manning has somehow been a better version of himself as he adjusted to his new body. He threw 37 touchdowns his first season with the Broncos, while spending more time in the training room and with his doctors than in the weight room as he worked to regain strength in his right triceps and waited for his nerve damage to improve.	• **The Broncos entered the 2014 season** as the defending AFC champions, hoping to compete for another Super Bowl run, following a 43–8 loss to the Seattle Seahawks in Super Bowl XLVIII. • Manning threw a total of 40 touchdown passes, but only four came in the last four games of the regular season and the playoffs.	• He threw 37 touchdowns his first **season with the Broncos**, while spending more time in the training room and with his doctors than in the weight room as he worked to regain strength in his right triceps and waited for his nerve damage to improve.	• no mention of 'why' in any related documents.

By Week 1 of the 2014 season, Manning had at least 1 touchdown pass in the 37 games he has played for the Broncos.
Prphr 1: By the kickoff of the 2014 season, Manning had achieved a touchdown pass in 37 of the contests he had featured in for the Broncos.
Prphr 2: At the onset of 2014 season Manning had at least one touchdown pass tallied in the 37 competitions participating by the Broncos.
Prphr 3: By week 1 of the 2014 season, Manning had tossed over one touchdown pass in the thirty seven contests he participated in for the Broncos.
Prphr 4: By the first week of the 2014 season, Manning had a minimum of one touchdown pass in all 37 matches he had played for the Broncos.
Prphr 5: When Week 1 of the 2014 season came around, Manning had attained 1 touchdown pass at least throughout the 37 games he had played for the Broncos.

Figure 14: Claims paraphrased using `text-davinci-003`

Figure 13: Claim: By Week 1 of the 2014 season , Manning had at least 1 touchdown pass in the 37 games he has played for the Broncos.

Who claims	What claims	When claims	Where claims	Why claims
No claim!	• **Q1**: ***What is controversial about city morgues music videos?*** Ans: Heavy use of drugs , violence , firearms , and nudity.	No claim!	No claim!	• **Q1**: ***Why are city morgue's music videos controversial?*** Ans: As they show heavy use of drugs , violence , firearms , and nudity.
? not verifiable	verified valid	? not verifiable	? not verifiable	verified valid
Evidence				
• no mention of 'who' in any related documents.	• City Morgue is an American hip hop group from New York, best known for their controversial music videos depicting the **heavy use of narcotics, violence, weaponry (mainly firearms), and nudity.**	• no mention of 'when' in any related documents.	• no mention of 'where' in any related documents.	• City Morgue is an American hip hop group from New York, best known for their controversial music videos depicting the **heavy use of narcotics, violence, weaponry (mainly firearms), and nudity.**

Figure 15: Claim: City Morgue s music videos are controversial as they show heavy use of drugs , violence , firearms , and nudity.

City Morgue s music videos are controversial as they show heavy use of drugs , violence , firearms , and nudity.
Prphr 1: City Morgue's song visuals are controversial as they present hefty utilization of drugs, savagery, firearms, and nakedness.
Prphr 2: City Morgue's music videos are seen as disputable due to their extensive portrayal of drug usage, brutality, weaponry, and nudity.
Prphr 3: City Morgue's music video content has caused debate for its graphically demonstrating of narcotics, brutality, firearms, and stark nudity.
Prphr 4: City Morgue has become notorious for the contentiousness of their music videos due to its frank exhibition of drugs, violence, guns, and nudity.
Prphr 5: City Morgue's music videos have been deemed controversial due to the inclusion of drugs, violence, guns, and nudity.

Figure 16: Claims paraphrased using `text-davinci-003`

Who claims	What claims	When claims	Where claims	Why claims
No claim!	• **Q1**: ***What movie was nominated for best animated feature and best original score?*** Ans: How to Train Your Dragon.	No claim!	• **Q1**: ***Where was how to train your dragon nominated for an academy award?*** Ans: At the 83rd Academy Awards.	No claim!
? not verifiable	verified valid	? not verifiable	verified valid	? not verifiable
Evidence				
• no mention of 'who' in any related documents	• **How to Train Your Dragon** premiered at the Gibson Amphitheater on March 21, 2010, and was released in the United States five days later on March 26. The film was a commercial success, earning nearly $500 million worldwide. It was widely acclaimed, being praised for its animation, voice acting, writing, musical score, and 3D sequences. It was nominated for the Academy Award for Best Animated Feature and Best Original Score at the 83rd Academy Awards, but lost to Toy Story 3 and The Social Network, respectively.	• no mention of 'when' in any related documents.	• How to Train Your Dragon premiered at the Gibson Amphitheater on March 21, 2010, and was released in the United States five days later on March 26. The film was a commercial success, earning nearly $500 million worldwide. It was widely acclaimed, being praised for its animation, voice acting, writing, musical score, and 3D sequences. It was nominated for the Academy Award for Best Animated Feature and Best Original Score at **the 83rd Academy Awards**, but lost to Toy Story 3 and The Social Network, respectively.	• no mention of 'why' in any related documents.

Figure 17: Claim: How to Train Your Dragon was nominated for the Academy Award for Best Animated Feature and Best Original Score at the 83rd Academy Awards.

How to Train Your Dragon was nominated for the Academy Award for Best Animated Feature and Best Original Score at the 83rd Academy Awards .
Prphr 1: The Academy Award was made to How to Train Your Dragon for Best Animated Feature and Best Original Score at the 83rd Academy Awards.
Prphr 2: How to Train Your Dragon earned a nomination for the Academy Award for Best Animated Feature and Best Original Score at the 83rd Academy Awards.
Prphr 3: How to Train Your Dragon got selected for the Academy Award for Best Animated Feature and Best Original Score at the 83rd Academy Awards.
Prphr 4: How to Train Your Dragon was put forward for the Academy Award for Best Animated Feature and Best Original Score at the 83rd Academy Awards.
Prphr 5: At the 83rd Academy Awards, the nomination of How to Train Your Dragon was bagged in the Best Animated Feature and Best Original Score categories.

Figure 18: Claims paraphrased using `text-davinci-003`

Who claims	What claims	When claims	Where claims	Why claims
• **Q1**: ***Who was sent off in the 42nd minute at manchester city?*** Ans: Medhi Benatia.	• No claim!	• **Q1**: ***When was medhi benatia sent off?*** Ans: In the 42nd minute.	• No claim!	• **Q1**: ***Why was medhi benatia sent off?*** Ans: For an infraction on Fernandinho.
verified valid	not verifiable	verified false	not verifiable	verified false
Evidence				
• On 17 September 2014, **Benatia** made his official debut for Bayern in a 1–0 home win against Manchester City, for the opening match of the 2014–15 UEFA Champions League season, where he played for 85 minutes, completing 93% of his passes. In the return match at Manchester City, he was sent off in the 20th minute for denying Sergio Agüero a clear goalscoring opportunity	• no mention of 'what' in any related documents.	• In the return match at Manchester City , he was sent off in the **20th minute** for denying Sergio Agüero a clear goalscoring opportunity ; the subsequent penalty was converted by Agüero and City went on to win 3–2.Benatia scored his first goal for Bayern on 13 December , opening the scoring in a 4–0 win at FC Augsburg with a header ; this result put his club 10 points clear at the top of the Bundesliga table.	• **In the return match at Manchester City** , he was sent off in the 20th minute for denying Sergio Agüero a clear goalscoring opportunity.	• In the return match at Manchester City , **he was sent off in the 20th minute for denying Sergio Agüero a clear goalscoring opportunity** ; the subsequent penalty was converted by Agüero and City went on to win 3–2.Benatia scored his first goal for Bayern on 13 December , opening the scoring in a 4–0 win at FC Augsburg with a header ; this result put his club 10 points clear at the top of the Bundesliga table.

Figure 19: Claim: In the return match at Manchester City , Medhi Benatia was sent off in the 42nd minute for an infraction on Fernandinho.

In the return match at Manchester City , Medhi Benatia was sent off in the 42nd minute for an infraction on Fernandinho.
Prphr 1: In the rematch conducted at Manchester City, Mehi Benatia was dismissed in the 42nd minute as he committed a foul towards Fernandinho.
Prphr 2: Back at Manchester City for the return game, Mehi Benatia was penalized with a red card in the 42nd minute for the infraction on Fernandinho.
Prphr 3: In the game held again at Manchester City, Medhi Benatia got his marching orders in the 42nd minute for a foul on Fernandinho.
Prphr 4: In the game held again in Manchester City, Medhi Benatia got a red card in the 42nd minute due to an infraction on Fernandinho.
Prphr 5: It was in Manchester City for the rematch when Medhi Benatia was shown the red card in the 42nd minute as a consequence of a grave infraction on Fernandinho.

Figure 20: Claims paraphrased using `text-davinci-003`

Who claims	What claims	When claims	Where claims	Why claims
• **Q1**: ***Who produced avengers assemble?*** Ans: The director of action movie Batman : Mask of the Phantasm.	• **Q1**: ***what was the name of the movie produced by batman : mask of the phantasm director?*** Ans: Avengers Assemble.	• **Q1**: ***When did avengers assemble premiere?*** Ans: On May 26 , 2013.	• **Q1**: ***Where did avengers assemble premiere?*** Ans: On Disney XD.	No claim!
verified valid	verified valid	verified valid	verified valid	not verifiable
Evidence				
• **Eric Radomsky** is one of the producers and directors of Avengers Assemble. He is also the Marvel Animation's Senior Vice President and Creative Director of Animation. He is perhaps best known as co-creator and co-producer of the Emmy award-winning Batman: Mask of the Phantasm.	• Eric Radomsky is one of the producers and directors of **Avengers Assemble**. He is also the Marvel Animation's Senior Vice President and Creative Director of Animation. He is perhaps best known as co-creator and co-producer of the Emmy award-winning Batman: Mask of the Phantasm.	• **M.O.D.O.K. Avengers Assemble** is an animated series, based on the fictional Marvel Comics superhero team the Avengers, which has been designed to capitalize on the success of The Avengers. Avengers Assemble premiered on May 26, 2013, on Disney XD.	• M.O.D.O.K. Avengers Assemble is an animated series, based on the fictional Marvel Comics superhero team the Avengers, which has been designed to capitalize on the success of The Avengers. Avengers Assemble premiered on May 26, 2013, on **Disney XD**.	• no mention of 'why' in any related documents.

Figure 21: Claim: The director of action movie Batman: Mask of the Phantasm, produced Avengers Assemble that premiered on Disney XD on May 26, 2013.

The director of action movie Batman: Mask of the Phantasm, produced Avengers Assemble that premiered on Disney XD on May 26, 2013 .
Prphr 1: The director of Batman: Mask of the Phantasm, which is an action flick, created Avengers Assemble and it made its premiere on Disney XD on May 26th, 2013.
Prphr 2: The director of the action-thriller Batman: Mask of the Phantasm authored Avengers Assemble premiering on the Disney XD portal on 26 May 2013.
Prphr 3: The director behind the action movie Batman: Mask of the Phantasm gave birth to Avengers Assemble viewed on Disney XD 26th May 2013.
Prphr 4: The Batman: Mask of the Phantasm director was also responsible for Avengers Assemble which debuted on Disney XD on 26/05/2013.
Prphr 5: The person who worked as the director for the action movie Batman: Mask of the Phantasm made Avengers Assemble and it was first aired on Disney XD on May 26th 2013.

Figure 22: Claims paraphrased using `text-davinci-003`

Who claims	What claims	When claims	Where claims	Why claims
• **Q1**: ***Who was benched during houston's game against texas tech?*** Ans: Allen.	No claim!	• **Q1**: ***when was allen benched?*** Ans: During Houstons game against Texas Tech.	No claim!	No claim!
verified valid	? not verifiable	verified valid	? not verifiable	? not verifiable
Evidence				
• **Kyle Allen** began last season as UH's starting quarterback, but he was benched in a loss to Texas Tech and only play briefly the remainder of the year.	• no mention of 'what' in any related documents.	• Kyle Allen had options to Stay at the University of Houston for another season, without the promise of ever seeing the football field again. But after a three turnover performance against Texas Tech on Sept. 23, Allen was benched and replaced by Kyle Postma who took over as the starter.	• no mention of 'where' in any related documents.	• no mention of 'why' in any related documents.

Figure 23: Claim: Allen was benched during Houston s game against Texas Tech.

Allen was benched during Houston s game against Texas Tech.
Prphr 1: Allen was removed from his position while Houston was facing Texas Tech.
Prphr 2: Allen was taken off the field during the Houston-Texas Tech match.
Prphr 3: Allen was put on the sideline during Houston's contest versus Texas Tech.
Prphr 4: Allen was out of the running during Houston's face of against Texas Tech.
Prphr 5: Allen was forbidden from playing during Houston's contest against Texas Tech.

Figure 24: Claims paraphrased using `text-davinci-003`

Who claims	What claims	When claims	Where claims	Why claims
• **Q1**: ***Who said?*** Ans: Kamala Harris.	• **Q1**: ***What did Kamala Harris say?*** Ans: "if you are going to be standing in that line for all those hours,you can't have any food."	No claim!	• **Q1**: ***Where are people supposed to stand?*** Ans: In line.	No claim!
verified false	verified false	? not verifiable	verified valid	? not verifiable
Evidence				
• Vice President Kamala Harris said that **state lawmakers have proposed hundreds of laws that will suppress or make it difficult for people to vote, and that one way state lawmakers have sought to curtail access to ballot is to cut off food or water to voters in line.**	• Vice President Kamala Harris said that **state lawmakers have proposed hundreds of laws that will suppress or make it difficult for people to vote, and that one way state lawmakers have sought to curtail access to ballot is to cut off food or water to voters in line.**	• no mention of 'when' in any related documents.	• Vice President Kamala Harris said that **state lawmakers have proposed hundreds of laws that will suppress or make it difficult for people to vote, and that one way state lawmakers have sought to curtail access to ballot is to cut off food or water to voters in line.**	• no mention of 'why' in any related documents.

Figure 25: Claim: Kamala Harris said that the new and proposed state laws on voting mean "if you are going to be standing in that line for all those hours, you can't have any food."

Kamala Harris said that the new and proposed state laws on voting mean "if you are going to be standing in that line for all those hours, you can't have any food."
Prphr 1: Kamala Harris expressed that the new and intended state regulations on voting mean "in case you are in the queue for all those hours, there is no eatables allowed."
Prphr 2: Kamala Harris spoke that the current and planned state legislations related to voting signify "if you are standing in that line for all that time, you cannot have any food."
Prphr 3: Kamala Harris highlighted that the recent and put forward state rules on voting mean that there is no food allowed while standing in line.
Prphr 4: Kamala Harris has commented on the new state laws on voting, proclaiming that people waiting in the long queue are not able to consume food.
Prphr 5: Kamala Harris mentioned that the state regulations being contended for voting have the stipulation that individuals who are standing in line for a prolonged period of time are not allowed to be eating.

Figure 26: Claims paraphrased using `text-davinci-003`

Who claims	What claims	When claims	Where claims	Why claims
No claim!	• **Q1**: ***What begin in the philippines?*** Ans: The start of coronavirus.	• **Q1**: ***when did the woman from wuhan arrive in manila?*** Ans: After traveling to Cebu City.	No claim!	No claim!
not verifiable	**verified valid**	**verified valid**	**? not verifiable**	**? not verifiable**
Evidence				
• no mention of 'who' in any related documents.	• Philippine health officials have confirmed the first case of the new coronavirus in the country.A 38-year-old Chinese woman, who arrived in the country from Wuhan, China, on Jan. 21, tested positive for **the novel coronavirus**, Health Secretary Francisco Duque told a news conference.	• A 38-year-old Chinese woman, who arrived in the country from Wuhan, China, on Jan. 21, tested positive for the novel coronavirus . DOH Epidemiology Bureau Director Ferchito Avelino said they are also looking at places where the woman stayed in Cebu and Dumaguete. He added that they are working to identify and quarantine employees at establishments who had close contact with the woman.	• no mention of 'where' in any related documents.	• no mention of 'why' in any related documents.

Figure 27: Claim: The start of coronavirus in the Philippines was a 38-year-old woman from Wuhan who arrived in Manila after traveling to Cebu City.

The start of coronavirus in the Philippines was a 38-year-old woman from Wuhan who arrived in Manila after traveling to Cebu City.
Prphr 1: The emergence of coronavirus in the Philippines was sparked by a 38-year-old female from Wuhan who made her way to Manila following her visit to Cebu City.
Prphr 2: The onset of coronavirus in the Philippines was initiated by a 38-year-old female from Wuhan who had visited Manila after traveling to Cebu City.
Prphr 3: The beginning of coronavirus in the Philippines was started by a 38-year-old female from Wuhan who moved to Manila after going to Cebu City.
Prphr 4: The initial appearance of the coronavirus in the Philippines came from a 38-year-old female from Wuhan who journeyed to Manila via Cebu City.
Prphr 5: The first time the coronavirus arrived in the Philippines was with a 38-year-old female from Wuhan that stopped by Manila after a trip to Cebu City.

Figure 28: Claims paraphrased using `text-davinci-003`

Who claims	What claims	When claims	Where claims	Why claims
• **Q1**: ***Who wrote the book series that robbie coltrane is based on?*** Ans: By J. K. Rowling.	• **Q1**: ***What is robbie coltrane known for?*** Ans: For his roles as a fictional character.	No claim!	No claim!	No claim!
verified valid	**verified valid**	**? not verifiable**	**? not verifiable**	**? not verifiable**
Evidence				
• Coltrane was widely known for starring in the "Harry Potter" franchise, based on the books by **J.K. Rowling**, alongside Daniel Radcliffe in the title role.	• Anthony Robert McMillan OBE (30 March 1950 – 14 October 2022), known professionally as Robbie Coltrane, was a Scottish actor and comedian. **He gained worldwide recognition in the 2000s for playing Rubeus Hagrid in the Harry Potter film series**.	• no mention of 'when' in any related documents.	• no mention of 'where' in any related documents.	• no mention of 'why' in any related documents.

Figure 29: Claim: Robbie Coltrane is known for his film roles as a fictional character based on a book series written by J. K. Rowling.

Robbie Coltrane is known for his film roles as a fictional character based on a book series written by J. K. Rowling.
Prphr 1: Robbie Coltrane is famed for his movie performances of a fictional character inspired by a set of books written by J. K. Rowling.
Prphr 2: Robbie Coltrane is renowned for his film parts as a fictional character originated from a book series composed by J. K. Rowling.
Prphr 3: Robbie Coltrane is well-known for his parts in films inspired by a book collection from J. K. Rowling.
Prphr 4: Robbie Coltrane rose to popularity because of the parts he played in films based off of the fictional work of J. K. Rowling.
Prphr 5: Robbie Coltrane is admired for his roles in pictures as a fictional character drawn from a collection of literature written by J. K. Rowling.

Figure 30: Claims paraphrased using `text-davinci-003`

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 9

☒ A2. Did you discuss any potential risks of your work?
This work doesn't have any risks

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1

☒ A4. Have you used AI writing assistants when working on this paper?
No, we didn't need it

B ☒ Did you use or create scientific artifacts?

Left blank.

☒ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.