Generating Literal and Implied Subquestions to Fact-check Complex Claims

Jifan Chen **Aniruddh Sriram** **Eunsol Choi** **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
jfchen@cs.utexas.edu

Abstract

Verifying political claims is a challenging task, as politicians can use various tactics to subtly misrepresent the facts for their agenda. Existing automatic fact-checking systems fall short here, and their predictions like "half-true" are not very useful in isolation, since it is unclear which parts of a claim are true or false. In this work, we focus on decomposing a complex claim into a comprehensive set of yes-no subquestions whose answers influence the veracity of the claim. We present CLAIMDECOMP, a dataset of decompositions for over 1000 claims. Given a claim and its verification paragraph written by fact-checkers, our trained annotators write subquestions covering both explicit propositions of the original claim and its implicit facets, such as additional political context that changes our view of the claim's veracity. We study whether state-of-the-art pretrained models can learn to generate such subquestions. Our experiments show that these models generate reasonable questions, but predicting implied subquestions based only on the claim (without consulting other evidence) remains challenging. Nevertheless, we show that predicted subquestions can help identify relevant evidence to fact-check the full claim and derive the veracity through their answers, suggesting that claim decomposition can be a useful piece of a fact-checking pipeline.[1]

1 Introduction

Despite a flurry of recent research on automated fact-checking (Wang, 2017; Rashkin et al., 2017; Volkova et al., 2017; Ferreira and Vlachos, 2016; Popat et al., 2017; Tschiatschek et al., 2018), we remain far from building reliable fact-checking systems (Nakov et al., 2021). This challenge motivated us to build more explainable models so the explanations can at least help a user interpret the results

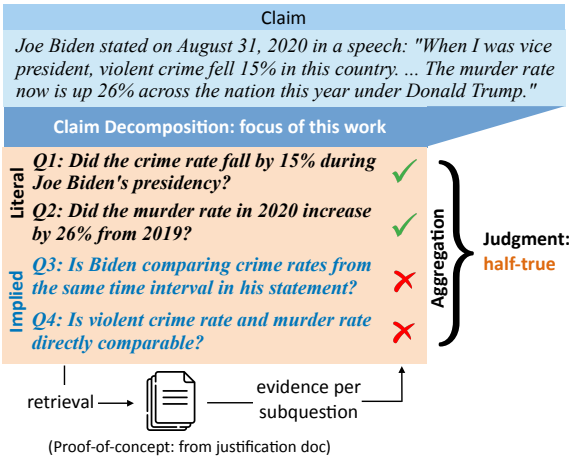


Figure 1: An example claim decomposition: the top two subquestions follow explicitly from the claim and the bottom two represent implicit reasoning needed to verify the claim. We can use the decomposed questions to retrieve relevant evidence (Section 6), and aggregate the decisions of the sub-questions to derive the final veracity of the claim (Section 5.3).

(Atanasova et al., 2020). However, such purely extractive explanations do not necessarily help users interpret a model's reasoning process. An ideal explanation should do what a human-written fact-check does: systematically dissect different parts of the claim and evaluate their veracity.

We take a step towards explainable fact-checking with a new approach and accompanying dataset, CLAIMDECOMP, of decomposed claims from PolitiFact. Annotators are presented with a claim *and* the justification paragraph written by expert fact-checkers, from which they annotate a set of yes-no subquestions that give rise to the justification. These subquestions involve checking both the explicit and implicit aspects of the claim (Figure 1).

Such a decomposition can play an important role in an interpretable fact verification system. First, the subquestions provide a comprehensive explanation of how the decision is made: in Figure 1, although the individual statistics mentioned by Biden

[1] We release our code and dataset: https://jifan-chen.github.io/ClaimDecomp

Claim: A Facebook post stated on January 31, 2021: "Nancy Pelosi bought $1.25 million in Tesla stock the day before Joe Biden signed an order "for all federal vehicles" to be electric."

Justification: An image shared on Facebook claims that Nancy Pelosi bought $1.25 million in Tesla stock the day before Biden signed an order for all federal vehicles to be electric, implying that she sought to profit from inside information about new government policies. The House speaker did report transactions involving Tesla stock, but the post misrepresented the purchases and Biden's policies to create the false impression that the transactions represented improper insider trading in Tesla shares.

Annotation: Question	Answer	Question Source	
Were the stock purchases improper insider trading?	No	Claim ○	Justification ●
Does the executive order Biden signed require all federal vehicles to be electric?	Unknown	Claim ●	Justification ○
Did Nancy Pelosi buy 1.25 million Tesla stock the day before Joe Biden signed an order about electric vehicles?	Unknown	Claim ●	Justification ○

Figure 2: An example of our annotation process. The annotators are instructed to write a set of subquestions, give binary answers to them, and attribute them to a source. If the answer cannot be decided from the justification paragraph, "Unknown" is also an option. The question is either based on the claim or justification, and the annotators also select the relevant parts (color-coded in the figure) on which the question is based.

are correct, they are from different time intervals and not directly comparable, which yields the final judgment of the claim as "half-true". We can estimate the veracity of a claim using the decisions of the subquestions (Section 5.3). Second, we show that decomposed subquestions allow us to retrieve more relevant paragraphs from the verification document than using the claim alone (Section 6), since some of the subquestions tackle implicit aspects of a claim. We do not build a full pipeline for fact verification in this paper, as there are other significant challenges this poses, including information which is not available online or which needs to be parsed out of statistical tables (Singh et al., 2021). Instead, we focus on showing how these decomposed questions can fit into a fact-checking pipeline through a series of proof-of-concept experiments.

Equipped with CLAIMDECOMP dataset, we train a model to generate decompositions of complex political claims. We experiment with pre-trained sequence-to-sequence models (Raffel et al., 2020), generating either a sequence of questions or a single question using nucleus sampling (Holtzman et al., 2020) over multiple rounds. This model can recover 58% of the subquestions, including some implicit subquestions. To summarize, we show that decomposing complex claims into subquestions can be learned with our dataset, and reasoning with such subquestions can lead improve evidence retrieval and judging the veracity of the whole claim.

2 Motivation and Task

Facing the complexities of real-world political claims, simply giving a final veracity to a claim often fails to be persuasive (Guo et al., 2022). To make the judgment of an automatic fact-checking system understandable, most previous work has focused on generating *justifications* for models' decisions. Popat et al. (2018); Shu et al. (2019); Lu and Li (2020) used attention weights of the models to highlight the most relevant parts of the evidence, but these only deal with explicit propositions of a claim. Ahmadi et al. (2019); Gad-Elrab et al. (2019) used logic-based systems to generate justifications, yet the systems are often based on existing knowledge graphs and are hard to adapt to complex real-world claims. Atanasova et al. (2020) treated the justification generation as a summarization problem in which they generate a justification paragraph according to some relevant evidence. Even so, it is hard to know which parts of the claim are true and which are not, and how the generated paragraph relates to the veracity.

What is missing in the literature is a better intermediate representation of the claim: with more complex claims, explaining the veracity of a whole claim at once becomes more challenging. Therefore, we focus on decomposing the claim into a **minimal** yet **comprehensive** set of yes-no subquestions, whose answers can be aggregated into an inherently explainable decision. As the decisions to the subquestions are explicit, it is easier for one to spot the discrepancies between the veracity and the intermediate decisions.

Claims and Justifications Our decomposition process is inspired by fact checking documents written by professional fact checkers. In the data we use from PolitiFact, each **claim** is paired with a **justification paragraph** (see Figure 2) which contains the most important factors on which the

Split	# unique claims	# tokens per claim	avg. # subquestions in single annotation	Answer %			Source %	
				Yes	No	Unknown	Justification	Claim
Train	800	33.4	2.7	48.9	45.3	5.8	83.6	16.4
Validation	200	33.8	2.7	48.3	44.8	6.9	79.0	21.0
Validation-sub	50	33.7	2.9	45.2	47.8	7.0	90.4	9.6
Test	200	33.2	2.7	45.8	43.1	11.1	92.1	7.9

Table 1: Statistics of the CLAIMDECOMP dataset. Each claim is annotated by two annotators, yielding a total of 6,555 subquestions. The second column blocks (Answer % and Source %) report the statistics at the subquestion level; Source % denotes the percentage of subquestions based on the text from the justification or the claim.

veracity made by the fact-checkers is based. Understanding *what questions are answered in this paragraph* will be the core task our annotators will undertake to create our dataset. However, we frame the claim decomposition task (in the next section) without regard to this justification document, as it is not available at test time.

Claim Decomposition Task We define the task of complex claim decomposition. Given a claim c and the context o of the claim (speaker, date, venue of the claim), the goal is to generate a set of N yes-no subquestions $\mathbf{q} = \{q_1, q_2, ...q_N\}$. The **set** of subquestions should have the following properties:

- **Comprehensiveness:** The questions should cover as many aspects of the claim as possible: the questions should be sufficient for someone to judge the veracity of the claim.
- **Conciseness:** The question set should be as minimal as is practical and not contain repeated questions asking about minor, correlated variants seeking the same information.

An individual subquestion should also exhibit:

- **Relevance:** The answer to subquestion should help a reader determine the veracity of the claim. Knowing an answer to a subquestion should change the reader's belief about the veracity of the original claim (Section 5.3).
- **Fluency / Clarity:** Each subquestion should be clear, fluent, and grammatically correct (Section 3).

We do not require subquestions to stand alone (Choi et al., 2021); they are instead interpreted with respect to the claim and its context.

Evaluation Metric We set the model to generate the target number of subquestions, which matches the number of subquestions in the reference, guaranteeing a concise subquestion set. Thus, we focus on measuring the other properties with reference-based evaluation. Specifically, given an annotated set of subquestions and an automatically predicted set of subquestions, we assess **recall**: how many subquestions in the reference set are covered by the generated question set? A subquestion in the reference set is considered as being recalled if it is **semantically equivalent** to one of the generated subquestions by models.[2] Our notion of equivalence is nuanced and contextual: for example, the following two subquestions are considered semantically equivalent: *"Is voting in person more secure than voting by mail?"* and *"Is there a greater risk of voting fraud with mail-in ballots?"*. We manually judge the question equivalence, as our experiments with automatic evaluation metrics did not yield reliable results (details in Appendix E).

3 Dataset Collection

Claim / Verification Document Collection We collect political claims and corresponding verification articles from PolitiFact.[3] Each article contains one justification paragraph (see Figure 2) which states the most important factors on which the veracity made by the fact-checkers is based. Understanding what questions are answered in this paragraph will be the core annotation task. Each claim is classified as one of six labels: *pants on fire* (most false), *false*, *barely true*, *half-true*, *mostly true*, and *true*. We collect the claims from top 50 PolitiFact pages for each label, resulting in a total of 6,859 claims.

A claim like *"Approximately 60,000 Canadians currently live undocumented in the USA."* hinges on checking a single statistic and is less likely to contain information beyond the surface form. Therefore, we mainly focus on studying complex claims

[2]There are cases where one generated question covers several reference questions, e.g., treating the whole claim as a question, in which case we only consider one of the reference questions to be recalled.

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

	ALL QS	MORE QS	FEWER QS
% of unmatched Qs	18.4	26.1	8.5

Table 2: Inter-annotator agreement assessed by the percentage of questions for which the semantics cannot be matched to the other annotator's set. We name the question set containing more questions as MORE QS and the other one as LESS QS. ALL QS is the average of MORE QS and LESS QS.

in this paper. To focus on complex claims, we filter claims with 3 or fewer verbs. We also filter out claims that do not have an associated justification paragraph. After the filtering, we get a subset consisting 1,494 complex claims.

Decomposition Annotation Process Given a claim paired with the justification written by the professional fact-checker on PolitiFact, we ask our annotators to reverse engineer the fact-checking process: generate yes-no questions which are answered in the justification. As shown in Figure 2, for each question, the annotators also (1) give the answer; (2) select the relevant text in the justification or claim that is used for the generation (if any). The annotators are instructed to cover as many of the assertions made in the claim as possible without being overly specific in their questions.

This process gives rise to both **literal questions**, which follow directly from the claim, and **implied questions**, which are not necessarily as easy to predict from the claim itself. These are not attributes labeled by the annotators, but instead labels the authors assign post-hoc (described in Section 5).

We recruit 8 workers with experience in literature or politics from the freelancing platform Upwork to conduct the annotation. Appendix A includes details about the hiring process, workflow, as well as instructions and the UI.

Dataset statistics and inter-annotator agreement Table 1 shows the statistics of our dataset. We collect two sets of annotations per claim to improve subquestion coverage. We collect a total of 6,555 subquestions for 1,200 claims. Most of the questions arise from the justification and most of the questions can be answered by the justification. In addition, we randomly sample 50 claims from the validation set for our human evaluation in the rest of this paper. We name this set **Validation-sub**.

Comparing sets of subquestions from different annotators is nontrivial: two annotators may choose different phrasings of individual questions and even different decompositions of the same claim that end up targeting the same pieces of information. Thus, we (the authors) manually compare two sets of annotations to judge inter-annotator agreement: given two sets of subquestions on the same claim, the task is to identify questions for which the semantics are not expressed by the other question *set*. If no questions are selected, it means that the two annotators show strong agreement on what should be captured in subquestions. Example annotations are shown in Appendix D.

We randomly sample 50 claims from our dataset and three of the authors conduct the annotation. The authors agree on this comparison task reasonably, with a Fleiss' Kappa (Fleiss, 1971) value of 0.52. The comparison results are shown in Table 2. On average, the semantics of 18.4% questions are not expressed by the other set. This demonstrates the **comprehensiveness** of our set of questions: only a small fraction is not captured by the other set, indicating that independent annotators are not easily coming up with distinct sets of questions. Because most questions are covered in the other set, we view the agreement as high. A simple heuristic to improve comprehensiveness further is to prefer the annotator who annotated more questions. If we consider the fraction of unmatched questions in the FEWER QS, we see this drops to 8.5%.[4] Through this manual examination, we also found that annotated questions are overall concise, fluent, clear, and grammatical.

4 Automatic Claim Decomposition

The goal is to generate a subquestion set $\mathbf{q}$ from the input claim c, the context o, and the target number of subquestions k.

Models We fine-tune a T5-3B (Raffel et al., 2020) model to automate the question generation process under two settings: QG-MULTIPLE and QG-NUCLEUS as shown in Figure 3. Both generation methods generate the same number of subquestions, equal to the number of subquestions generated by an annotator.

QG-MULTIPLE We learn a model $P(\mathbf{q} \mid c, o)$ to place a distribution over sets of subquestions given the claim and output. The annotated questions are

[4] Merging two annotations results in many duplicate questions and deduplicating these without another round of adjudication is cognitively intensive. We opted not to do this due to the effectiveness of simply taking the larger set of questions.

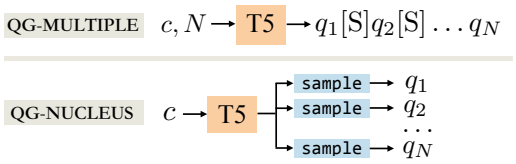


Figure 3: Illustration of our two question generators. QG-MULTIPLE generates all questions as a sequence while QG-NUCLEUS generates one question at a time through multiple samples.

Model	R-all	R-literal	R-implied
QG-MULTIPLE	0.58	0.74	0.18
QG-NUCLEUS	0.43	0.59	0.11
QG-MULTIPLE-JUSTIFY	0.81	0.95	0.50
QG-NUCLEUS-JUSTIFY	0.52	0.72	0.18

Table 3: Human evaluation results on the Validation-sub set (N=146). R-all denotes the recall for all questions; R-literal and R-implied denotes the recall for the literal questions and the implied questions respectively.

concatenated by their annotation order to construct the output.

QG-NUCLEUS We learn a model $P(q \mid c, o)$ to place a distribution over single subquestions given the claim and output. For training, each annotated subquestion is paired with the claim to form a *distinct* input-output pair. At inference, we use nucleus sampling to generate questions. See Appendix F for training details.

We also train these generators in an oracle setting where the justification paragraph is appended to the claim to understand how well the question generator does with more information. We denote the two oracle models as QG-MULTIPLE-VERIFY and QG-NUCLEUS-VERIFY respectively.

Results All models are trained on the training portion of our dataset and evaluated on the Validation-sub set. One of the authors evaluated the recall of each annotated subquestion in the generated subquestion set. The results are shown in Table 3. We observe that **most of the literal questions can be generated while only a few of the implied questions can be recovered.** Generating multiple questions as a single sequence (QG-MULTIPLE) is more effective than sampling multiple questions (QG-NUCLEUS). Many questions generated from QG-NUCLEUS are often slightly different but share the same semantics. We see that more than 70% of the literal questions and 18% of the implied questions can be generated by the best

Question Type	# Questions	R1-P	R2-P	RL-P
Literal	2.15	0.56	0.30	0.47
Implied	1.02	0.28	0.09	0.22

Table 4: Number of questions of each type per claim and their lexical overlap with the claim measured by ROUGE-1, ROUGE-2, and ROUGE-L precision (how many n-grams in the question are also in the claim).

QG-MULTIPLE model. By examining the generated implied questions, we find that most of them belong to the **domain knowledge** category in Section 5.

Some questions could be better generated if related evidence were retrieved first, especially for questions of the **context** category (Section 5). The QG-MULTIPLE-JUSTIFY model can recover most of the literal questions and half of the implied questions. Although this is an oracle setting, it shows that when given proper information about the claim, the T5 model can achieve much better performance. We discuss this retrieval step more in Section 9.

Qualitative Analysis While our annotated subquestion sets cover most relevant aspects of the claim, we find some generated questions are good subquestions that are missing in our annotated set, though less important. For example, for our introduction example shown in Figure 1, the QG-NUCLEUS model generates the question "*Is Trump responsible for the increased murder rate?*" Using the question generation model in collaboration with humans might be a promising direction for more comprehensive claim decomposition. See Appendix H for more examples.

5 Analyzing Decomposition Annotations

In this section, we study the characteristics of the annotated questions. We aim to answer: (1) How many of the questions address implicit facets of the claim, and what are the characteristics of these? (2) How do our questions differ from previous work on question generation for fact checking (Fan et al., 2020)? (3) Can we aggregate subquestion judgments for the final claim judgment?

5.1 Subquestion Type Analysis

We (the authors) manually categorize 285 subquestions from 100 claims in the development set into two disjoint sets: *literal* and *implied*, where *literal* questions are derived from the surface information of the claim – whether a question can be posed

Domain knowledge (38.8%)	**Claim**: "When President Obama was elected, the market crashed … Trump was up 9%, President Obama was down 14.8% and President Bush was down almost 4%. There is an instant reaction on Wall Street." **Question**: Did Obama cause the stock market crash when he was elected? (**Domain knowledge of whether the stock market is correlated with the election.**)
Context (37.6%)	**Claim**: With voting by mail, "you get thousands and thousands of people … signing ballots all over the place." **Question**: Is there a greater risk of voting fraud with mail-in ballots? (**Need to know the background that the claim is about the potential risks of mail-in ballots.**)
Implicit meaning (16.5%)	**Claim**: Nancy Pelosi bought $1.25 million in Tesla stock the day before Joe Biden signed an order "for all federal vehicles" to be electric. **Question**: Were the stock purchases improper insider trading? (**The claim implies this purchase is insider trading.**)
Statistical rigor (7.1%)	**Claim:** "No other country witnesses the number of gun deaths that we do here in the U.S., and it's not even close." **Question:** Is the United States the country with the the highest percentage of gun deaths? (**Highest number of gun deaths does not entail highest percentage of gun deaths.)**

Figure 4: Four types of reasoning needed to address subquestions with their proportion (left column) and examples (right column). It shows that a high proportion of the questions need either domain knowledge or related context.

by only given the claim, and *implied* questions are those that need extra knowledge in order to pose.

Table 4 shows basic statistics about these sets, including the average number of subquestions for each claim and lexical overlap between subquestions and the base claims, evaluated with ROUGE precision, as one subquestion can be a subsequence of the original claim. On average, each claim contains one implied question which represents the deeper meaning of the claim. These implied questions overlap less with the claim.

We further manually categorize the implied questions into the following four categories, reflecting what kind of knowledge is needed to pose them (examples in Figure 4). Two authors conduct the analysis over 50 examples and the annotations agree with a Cohen's Kappa (Cohen, 1960) score of 0.74.

Domain knowledge The subquestion seeks domain-specific knowledge, for example asking about further steps of a legal or political process.

Context The subquestion involves knowing that broader context is relevant, such as whether something is broadly common or the background of the claim (political affiliation of the politician, history of the events stated in the claim, etc).

Implicit meaning The subquestion involves unpacking the implicit meaning of the claim, specifically anchored to what the speaker's intent was.

Statistical rigor The subquestion involves checking over-claimed or over-generalized statistics (e.g., the highest raw count is not the highest per capita).

Most of the implied subquestions require either domain knowledge or context about the claim, reflecting the challenges behind automatically generating such questions.

	mean	std	# examples
QABriefs (Fan et al., 2020)	2.88	1.20	210
Ours	3.60	1.19	210
p-value	≤ 0.0001		
mean diff	0.72		
95% CI	0.48 - 0.97		

Table 5: Results from user study on helpfulness (rated 1-5) of a set of generated subquestions for claim verification. We conduct a t-test over the collected scores.

5.2 Comparison to QABriefs

Our work is closely related to the QABriefs dataset (Fan et al., 2020), where they also ask annotators to write questions to reconstruct the process taken by professional fact-checkers provided the claim and its verification document.

While sharing similar motivation, we use a significantly different annotation process than theirs, resulting in qualitatively different sets of questions as shown in Figure 5. We notice: (1) Their questions are less comprehensive, often missing important aspects of the claim. (2) Their questions are broader and less focused on the claim. We instructed annotators to provide the **source** of the annotated subquestions from either claim or verification document. For example, questions like "*What are Payday lenders?*" in the figure will not appear in our dataset as the justification paragraph does not address such question. Fan et al. (2020) dissuaded annotators from providing binary questions; instead, they gather answers to their subquestions after the questions are collected. We focus on binary questions whose verification could help verification of the full claim. See Appendix I for more examples of the comparison.

Claim: The group With Honor stated on September 10, 2018 in a TV ad: Kentucky Rep. Andy Barr "would let shady payday lenders take advantage of our troops" and that he took "$36,550 from payday lenders."

CLAIMDECOMP

1. Has Barr received $36,550 from payday lenders?
2. Did Barr vote for legislation that would weaken restrictions for payday lenders?
3. Are there any protections for service members using payday lending services?
4. Has Barr's voting record directly affected protection for veterans against payday lenders?

Fan et al. (2020)

1. What are Payday lenders?
   helpful background but not precisely about claim
2. What's the maximum amount you can get from payday lenders?
   useful context but not directly about claim
3. What percentage of US troops use a payday lender?
   useful context but not directly about claim

Figure 5: Comparison between our decomposed questions with QABriefs (Fan et al., 2020). In general, our decomposed questions are more comprehensive and relevant to the original claim.

	Macro-F1	Micro-F1	MAE
Question aggregation	0.30	0.29	1.05
Question aggregation*	0.46	0.45	0.73
Random (label dist)	0.16	0.18	1.68
Most frequent	0.06	0.23	1.31

Table 6: Claim classification performance of our question aggregation baseline vs. several baselines on the development set. MAE denotes mean absolute error.

User Study To better quantify the difference, we also conduct a user study in which we ask an annotator to rate how useful a set of questions (without answers) are to determine the veracity of a claim. On 42 claims annotated by both approaches, annotators score sets of subquestions on a Likert scale from 1 to 5, where 1 denotes that knowing the answers to the questions does not help at all and 5 denotes that they can accurately judge the claim once they know the answer. We recruit annotators from MTurk. We collect 5-way annotation for each example and conduct the t-test over the results. The details can be found in Appendix C.

Table 5 reports the user study results. Our questions achieve a significantly higher relevance score compared to questions from QABriefs. This indicates that we can potentially derive the veracity of the claim from our decomposed questions since they are binary and highly relevant to the claim.

5.3 Deriving the Veracity of Claims from Decomposed Questions

Is the veracity of a claim sum of its parts? We estimate whether answers to subquestions can be used to determine the veracity of the claim.

We predict a veracity score $\hat{v} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[a_i = 1]$ equal to the fraction of subquestions with yes answers. We can map this to the discrete 6-label scale by associating the labels *pants on fire*, *false*, *barely true*, *half true*, *mostly true*, and *true* with the intervals $[0, \frac{1}{6}), [\frac{1}{6}, \frac{2}{6}), [\frac{2}{6}, \frac{3}{6}), [\frac{3}{6}, \frac{4}{6}), [\frac{4}{6}, \frac{5}{6}), [\frac{5}{6}, 1]$, respectively. We call this method **question aggregation**. We use the 50 claims and the corresponding questions from the **Validation-sub** set for evaluation. We also establish the upper bound (**question aggregation***) for this heuristic by having one of the authors remove unrelated questions. On average, 0.3 questions are removed per claim.

Table 6 compares our heuristics with simple baselines (random assignment and most frequent class assignment). Our heuristic easily outperforms the baselines, with the predicted label on average is only shifted by one label, e.g., *mostly true* vs. *true*. This demonstrates the potential of building a more complex model to aggregate subquestion-answer sets, which we leave as a future direction.

Our simple aggregation suffers in the following cases: (1) The subquestions are not equal in importance. The first example in Figure 4 contains two yes subquestions and two no subquestions, and our aggregation yields *half-true* label, differing from gold label *barely-true*. (2) Not all questions are relevant. As indicated by **question aggregation***, we are able to achieve better performance after removing unrelated questions. (3) In few cases, the answer to a question could inversely correlate with the veracity of a claim. For example, the claim states "Person X implied Y" and the question asks "Did person X not imply Y?" We think all of the cases can be potentially fixed by stronger models. For example, a question salience model can mitigate (1) and (2), and promotes researches about understanding core arguments of a complex claim. We leave this as future work.

	per subquestion	per example (claim)
avg # of paras	12.4	12.4
% of context	87.6	68.8
% of support	5.4	12.0
% of refute	8.0	19.2
Fleiss Kappa	0.42	0.42

Table 7: Evidence paragraph retrieval data statistics on Validation-sub dataset (50 claims).

6 Evidence Retrieval with Decomposition

Lastly, we explore using claim decomposition for retrieving evidence paragraphs to verify claims. Retrieval from the web to check claims is an extremely hard problem (Singh et al., 2021). We instead explore a simplified proof-of-concept setting: retrieving relevant paragraphs from the full justification document. These articles are lengthy, containing an average of 12 paragraphs, and with distractors due to entity and concept overlap with the claims.

We aim to show two advantages of using the decomposed questions: (1) The implied questions contain information helpful to retrieve evidence beyond the lexical information of the claim. (2) We can convert the subquestions to statements and treat them as hypotheses to apply the off-the-shelf NLI models to retrieve evidence that entails such hypotheses (Chen et al., 2021).

Evidence Paragraph Collection We first collect human annotation to identify relevant evidence paragraphs. Given the full PolitiFact verification article consisting of m paragraphs $\mathbf{p} = (p_1, \ldots, p_m)$ and a subquestion, annotators find paragraphs relevant to the subquestion. As this requires careful document-level reading, we hire three undergraduate linguistics students as annotators. We use the 50 claims from the Validation-sub set and present the annotators with the subquestions and the articles. For each subquestion, for each paragraph in the article, we ask the annotators to choose whether it served as context to the subquestion or whether it supports/refutes the subquestion. The statistics and inter-annotator agreement is shown in Table 7. Out of 12.4 paragraphs on average, 3-4 paragraphs were directly relevant to the claim and the rest of paragraphs mostly provide context.

Experimental Setup We experiment with three off-the-shelf RoBERTa-based (Liu et al., 2019) NLI models trained on three different datasets: MNLI (Williams et al., 2018), NQ-NLI (Chen et al., 2021), and DocNLI (Yin et al., 2021). We compare the performance of NLI models with random, BM25, and human baselines.

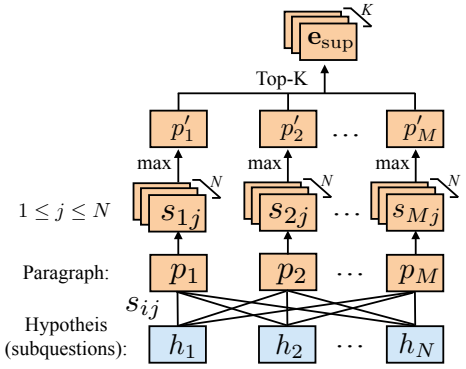


Figure 6: Illustration of evidence paragraph retrieval process. The notations corresponds to our descriptions in Section 6. K is a hyperparameter controlling the number of passages to retrieve.

Model	Decomposed claim predicted	Decomposed claim gold	Original claim
MNLI	41.0	48.8	35.2
NQ-NLI	38.8	34.5	40.9
DocNLI	**44.7**	**59.6**	36.9
BM25	36.2	47.5	39.2

Table 8: Evidence retrieval performance (F1 score) with the decomposed claims (from predicted and annotated (gold) subquestions) and the original claim on the Validation-sub set. A random baseline achieves 24.9 F1 and human annotators achieve 69.0 F1.

We first convert the corresponding subquestions $\mathbf{q} = q_1, ..., q_N$ of claim c to a set of statements $\mathbf{h} = h_1, ..., h_n$ using GPT-3 (Brown et al., 2020).[5] We find that with only 10 examples as demonstration, GPT-3 can perform the conversion quite well (with an error rate less than 5%). For more information about the prompt see Appendix G.

To retrieve the evidence that **supports** the statements, we treat the statements as hypotheses and the paragraphs in the article as premises. We feed them into an NLI model to compute the score associated with the "entailment" class for every premise and hypothesis pair. Here, the score for paragraph p_i and hypothesis h_j is defined as the output probability $s_{ij} = P(\text{Entailment} \mid p_i, h_j)$. We then select as evidence the top k paragraphs by score across all subquestions: for paragraph p_i, we define $p'_i = \max(\{s_{ij} \mid 1 \le j \le N\})$, which denotes for each hypothesis from 1 to N that the jth hypothesis h_j achieves the highest score with p_i. Then $\mathbf{e_{sup}} = \{p_i \mid i \in \text{Top-K}(\{p'_1, ..., p'_M\})\}$. We

[5] We release the automatically converted statements and the negations for all of the subquestions in the published dataset.

set k to be the number of the paragraphs that are annotated with either support or refute. Figure 6 describes this approach.

To retrieve the evidence that **refutes** the statements, we follow the same process, but with the negated hypotheses set $\mathbf{h}$ generated by GPT3. (Note that our NLI models trained on NQ-NLI and DocNLI only have two classes, entailed and not entailed, and not entailed is not a sufficient basis for retrieval.) The final evidence set is obtained by merging the evidence from the *support* and *refute* set. This is achieved by removing duplicates then taking Top-K paragraphs according to the scores.

BM25 baseline model uses retrieval score instead of NLI score. **The random baseline** randomly assign support, refute, neutral labels to paragraphs based on the paragraph label distribution in Table 7. **Human performance** is computed by selecting one of the three annotators and comparing their annotations with the other two (we randomly pick one annotator if they do not agree), taking the average over all three annotators. This is not directly comparable to the annotations for the other techniques as the gold labels are slightly different.

Results The results are shown in Table 8. We see that **the decomposed questions are effective to retrieve the evidence.** By aggregating evidence from the subquestions, both BM25 and the NLI models can do better than using the claim alone, except for the case of using DocNLI, and BM25 with the predicted decomposition. The best model with gold annotations (59.6) is close to human performance (69.0) in this limited setting, indicating that the detailed and implied information in decomposed questions can help gathering evidence beyond the surface level of the claim.

DocNLI outperforms BM25 on both the annotated decomposition and the predicted decomposition. This demonstrates the potential of using the NLI models to aid the evidence retrieval in the wild, although they must be combined with decomposition to yield good results.

7 Related Work

Fact-checking Vlachos and Riedel (2014) proposed to decompose the fact-checking process into three components: identifying check-worthy claims, retrieving evidence, and producing verdicts. Various datasets have been proposed, including human-generated claims based on Wikepedia (Thorne et al., 2018; Chen et al., 2019; Jiang et al., 2020; Schuster et al., 2021; Aly et al., 2021), real-world political claims (Wang, 2017; Alhindi et al., 2018; Augenstein et al., 2019; Ostrowski et al., 2021; Gupta and Srikumar, 2021), and science claims (Wadden et al., 2020; Saakyan et al., 2021). Our dataset focuses on real-world political claims, particularly more complex claims than past work which necessitate the use of decompositions.

Our implied subquestions go beyond what is mentioned in the claim, asking the intention and political agenda of the speaker. Gabriel et al. (2022) study such implications by gathering expected readers' reactions and writers' intentions towards news headlines, including fake news headlines.

To produce verdicts of the claims, other work generates explanations for models' predictions. Popat et al. (2017, 2018); Shu et al. (2019); Yang et al. (2019); Lu and Li (2020) presented attention-based explanations; Gad-Elrab et al. (2019); Ahmadi et al. (2019) used logic-based systems, and Atanasova et al. (2020); Kotonya and Toni (2020) modeled the explanation generation as a summarization task. Combining answers to the decomposed questions in our work can form an explicit explanation of the answer.

Question Generation Our work also relates to question generation (QG) (Du et al., 2017), which has been applied to augment data for QA models (Duan et al., 2017; Sachan and Xing, 2018; Alberti et al., 2019), evaluate factual consistency of summaries (Wang et al., 2020; Durmus et al., 2020; Kamoi et al., 2022), identify semantic relations (He et al., 2015; Klein et al., 2020; Pyatkin et al., 2020), and identify useful missing information in a given context (clarification) (Rao and Daumé III, 2018; Shwartz et al., 2020; Majumder et al., 2021). Our work is most similar to QABriefs (Fan et al., 2020), but differs from theirs in two ways: (1) We generate yes-no questions directly related to checking the veracity of the claim. (2) Our questions are more comprehensive and precise.

8 Conclusion

We present a dataset containing more than 1,000 real-world complex political claims with their decompositions in question form. With the decompositions, we are able to check the explicit and implicit arguments made in the claims. We also show the decompositions can play an important role in both evidence retrieval and veracity composition of an explainable fact-checking system.

9 Limitations

Interaction of retrieval and decomposition The evidence retrieval performance depends on the quality of the decomposed questions (compare our results on generated questions to those on annotated questions in Section 6). Yet, generating high-quality questions requires relevant evidence context. These two modules cannot be strictly pipelined and we envision that in future work, they will need to interact in an iterative fashion. For example, we could address this with a human-in-the-loop approach. First, retrieve some context passages with the claim to verify as a query, possibly focused on the background of the claim and the person who made the claim. This retrieval can be done by a system or a fact-checker. Then, we use context passages to retrain the QG model with the annotations we have and the fact-checker can make a judgment about those questions, adding new questions if the generated questions do not cover the whole claim. We envision that such a process can make fact-checking easier while providing data to train the retrieval and QG models.

Difficulty of automatic question comparison As discussed in section 4, automatic metrics to evaluate our set of generated questions do not align well with human judgments. Current automatic metrics are not sensitive enough to minor changes that could lead to different semantics for a question. For example, changing *"Are **all** students in Georgia required to attend chronically failing schools?"* to *"Are students in Georgia required to attend chronically failing schools?"* yields two questions that draw an important contrast. However, we will get an extremely high BERTScore (0.99) and ROUGE-L score (0.95) between the two questions. Evaluating question similarity without considering how the questions will be used is challenging, since we do not know what minor distinctions in questions may be important. We suggest measuring the quality of the generated questions on some downstream tasks, e.g., evidence retrieval.

General difficulty of the task We have not yet built a full pipeline for fact-checking in the true real-world scenario. Instead, we envision our proposed question decomposition as an important step of such a pipeline, where we can use the candidate decompositions to retrieve deeper information and verify or refute each subquestion, then compose the results of the subquestions into an inherently explainable decision. In this paper, we have shown that the decomposed questions can help the retriever in a clean setting. But retrieving evidence in the wild is extremely hard since some statistics are not accessible through IR and not all available information is trustworthy (Singh et al., 2021), which are issues beyond the scope of this paper. Through the **Question Aggregation** probing, we also show the potential of composing the veracity of claims through the decisions from the decomposed questions. The proposed dataset opens a door to study the core argument of a complex claim.

Domain limitations and lack of representation The dataset we collected only consists of English political claims from the PolitiFact website. These claims are US-centric and largely focused on politics; hence, our results should be interpreted with respect to these domain limitations.

Broader impact: opportunities and risks of deployment Automated fact checking can help prevent the propagation of misinformation and has great potential to bring value to society. However, we should proceed with caution as the output of a fact-checking system—the veracity of a claim—could alter users' views toward someone or something. Given this responsibility, we view it as crucial to develop explainable fact-checking systems which inform the users which parts of the claim are supported, which parts are not, and what evidence supports these judgments. In this way, even if the system makes a mistake, users will be able to check the evidence and draw the conclusion themselves.

Although we do not present a full fact-checking system here, we believe our dataset and study can help pave the way towards building more explainable systems. By introducing this claim decomposition task and the dataset, we will enable the community to further study the different aspects of real-world complex claims, especially the implied arguments behind the surface information.

Acknowledgments

This dataset was funded by Good Systems,[6] a UT Austin Grand Challenge to develop responsible AI technologies, as well as NSF Grant IIS-1814522, NSF CAREER Award IIS-2145280, and gifts from Salesforce and Adobe. We would like to thank

[6] https://goodsystems.utexas.edu/

our annotators from Upwork and UT Austin: Andrew Baldino, Scarlett Boyer, Catherine Coursen, Samantha Dewitt, Abby Mortimer, E Lee Riles, Keegan Shults, Justin Ward, Meona Khetrapal, Misty Peng, Payton Wages, and Maanasa V Darisi.

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. In *Conference on Truth and Trust Online*.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

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.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

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.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations*.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *ACL*.

Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 87–95.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, and Greg Durrett. 2022. Shortcomings of Question Answering Based Factuality Frameworks for Error Localization. In *arXiv*.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on 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.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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. In *arXiv*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*.

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)*.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012. International World Wide Web Conferences Steering Committee.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. 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 629–640, New Orleans, Louisiana. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Prakhar Singh, Anubrata Das, Junyi Jessy Li, and Matthew Lease. 2021. The Case for Claim Difficulty Assessment in Automatic Fact Checking. *arXiv ePrint 2109.09689*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez-Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *The Web Conference, Alternate Track on Journalism, Misinformation, and Fact-checking*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

647–653, Vancouver, Canada. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*, pages 3600–3604.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*.

A Question Annotation Workflow

A.1 Workflow

Tracing the thought process of professional fact-checkers requires careful reading. Thus, instead of using crowdsourcing platforms with limited quality control, we recruit 8 workers with experience in literature or politics from the freelancing platform Upwork.[7] We pay the annotators $1.75 per claim, which translates to around $30/hour.[8] Each annotator labeled an initial batch of articles and we provided feedback on their annotation. We communicated with annotators during the process.

We posted a job advertisement including the description and the payment plan of our task on the Upwork platform. In total 14 workers applied for the position. We first conducted an initial qualification round in which we released an initial batch of 15 documents for the annotators to complete, for which we paid $35. This initial batch was used to judge how suitable the annotators are for this task. We reviewed the annotations and give detailed feedback to each annotator for every claim along with our suggested annotation for reference. We selected annotators whose annotation met our qualifications to continue to the next round. In the initial round, we selected 8 out of the 14 annotators who applied.

After the initial round, we released new example batches to the annotators on a weekly basis. Each batch contained 100 examples for which we paid $175. The hired annotators were required to complete at least one batch per week and they could do up to 2 batches per week.[9]

A.2 Annotation Interface

The interface of the main question decomposition task is shown in Figure 7.

B Evidence Annotation Interface

The interface to annotate the supporting/refuting evidence described in section 6 is shown in Figure 8.

[7] https://www.upwork.com/

[8] We asked the workers to report their speeds at the end of the task and found their actual hourly rates ranged between $18 and $50 per hour.

[9] We intentionally limited this to avoid having a single annotator annotate a large portion of the examples.

C User Study Interface

The annotation interface of our user study conducted in Section 5.2 is shown in Figure 9.

D Inter-annotator Agreement

Two examples of our inter-annotator agreement assessment are shown in Figure 10. In the first example, we treat Q3 of annotator A as not covered by annotator B. It is a weaker version of Q2 but not mentioned by annotator B. Q4 of annotator A has similar semantics as Q3 of annotator B so we do not mark it.

E Automatic Claim Decomposition Evaluation

For the evaluation in Section 4, we also explored an automated method for assessing whether generated questions match ground truth ones. We aim to define a metric $m(\mathbf{q}, \hat{\mathbf{q}})$ that compares the two sets of generated questions. However, we lack good off-the-shelf methods for comparing sets of strings like this. Instead, we rely on existing scoring functions that can compare single strings, like ROUGE and BERTScore (Zhang et al., 2019).

Following other alignment-based methods like SummaC (Laban et al., 2022) for summarization factuality, we view these metrics as: $m(\mathbf{q}, \hat{\mathbf{q}}) = \text{argmax}_{\mathbf{a}} \sum s(q_{a_i}, \hat{q}_i)$, where $\mathbf{a}$ is an alignment variable. This problem can be viewed as finding the maximum-weight matching in a bipartite graph. We use the Hungarian algorithm (Kuhn, 1955) to compute this alignment and we take the mean of max matching as the result. The results are shown in Table 9.

The automatic metrics are not well aligned with the human judgments. We see that the Pearson coefficient between human judgments and the automatic metrics ranges from 0.42–0.54 and 0.21–0.45 for QG-MULTIPLE and QG-NUCLEUS respectively. The large instability of the Pearson coefficient indicates that the automatic evaluation may not accurately reflect the quality of the generated questions. Therefore, evaluating the generated questions on downstream tasks could be more accurate, hence why we also study evidence retrieval.

F Training Details for Question Generation

For QG-MULTIPLE, each instance of the input and the output are constructed according to the template:

$$N\text{[S]}c\text{[S]} \longrightarrow \text{q}_1\text{[S]}\text{q}_2\text{[S]}, ..., \text{q}_N$$

where [S] denotes the separator token of the T5 model. N denotes the number of questions to generate; we introduce this into the input to serve as a control variable and set it to match the number of annotated questions during training. c denotes the claim and q_i denotes the ith annotated question and we do not assume a specific order for the questions.

The model is trained using the seq2seq framework of Hugging Face (Wolf et al., 2020). The max sequence length for input and output is set to 128 and 256 respectively. The batch size is set to 8 and we use DeepSpeed for memory optimization (Rasley et al., 2020). We train the model on our training set for 20 epochs with AdamW (Loshchilov and Hutter, 2018) optimizer and an initial learning rate set to 3e-5.

At inference, we use beam search with beam size set to 5. We prepend the number of questions to generate (N) at the start of the claim in the input.

For QG-NUCLEUS, we construct multiple input-output instances $(c \rightarrow q_i)$ for each claim, where q_i denotes the ith decomposed question of claim c. The max sequence length for input and output are both set to 128. The batch size is set to 16 and we use DeepSpeed for memory optimization. We train the model on our training set for 10 epochs with AdamW optimizer and an initial learning rate set to 3e-5.

We expect this model to place a flatter distribution over the output space, assigning many possible questions high weight due to the training data including multiple outputs for the same input. At inference, we use nucleus sampling (Holtzman et al., 2019) in which p is set to 0.95 together with top-k sampling (Fan et al., 2018) in which k is set to 50 to generate questions. We filter out the duplicates (exact string match) in the sampled questions set.

G GPT-3 for Question Conversion

Given a question, we let GPT-3 generate its declarative form as well as the negated form of the statement. We achieve this by separating them using "|" in the prompt. One advantage of using GPT-3 is that it can easily generate natural sentences. For example, for question *"Are any votes illegally counted in the election?"*, GPT-3 generates the statement and its negation as *"Some votes were illegally counted in the election."* and

Model	Rouge-1 (P)	Rouge-2 (P)	Rouge-l (P)	Bert-score (P)
QG-MULTIPLE	0.44 (0.54)	0.23 (0.47)	0.40 (0.53)	0.92 (0.42)
QG-NUCLEUS	0.39 (0.32)	0.18 (0.21)	0.36 (0.32)	0.91 (0.45)
QG-MULTIPLE-JUSTIFY	0.54 (0.36)	0.38 (0.35)	0.52 (0.37)	0.93 (0.35)
QG-NUCLEUS-JUSTIFY	0.41 (0.25)	0.20 (0.35)	0.37 (0.30)	0.91 (0.41)

Table 9: Automatic evaluation results on the development set. Here, (P) denotes the Pearson correlation coefficient between the automatic metric and recall-all. -JUSTIFY denotes training the question generator by concatenating the claim and the justification paragraph as the input.

"No votes were illegally counted in the election.". A demonstration of the prompt we used for the question conversion is shown as follows:

Question: Are unemployment rates for African Americans and Hispanics low today?
Statement: Unemployment rates for African Americans and Hispanics are low today. | Unemployment rates for African Americans and Hispanics are not low today.

Question: Were 1700 mail-in ballots investigated for fraud in Texas?
Statement: 1700 mail-in ballots were investigated for fraud in Texas | 1700 mail-in ballots were not investigated for fraud in Texas

Question: Is Wisconsin guaranteeing Foxconn nearly $3 billion?
Statement: Wisconsin guarantees Foxconn nearly $3 billion. | Wisconsin does not guarantee Foxconn nearly $3 billion.

Question: Will changes in this law raise taxes for anyone?
Statement: The changes in this law will raise taxes for someone. | The changes in this law will raise taxes for no one.

Question: Has Donnelly directly sponsored any of these legislative proposals since becoming a senator?
Statement: Donnelly directly sponsored some of these legislative proposals since becoming a senator. | Donnelly directly sponsored none of these legislative proposals since becoming a senator.

...

Question: INPUT-QUESTION
Statement: MODEL-OUTPUT

H Qualitative Analysis of Generated Questions

Table 12 includes more examples where the generated questions do not match the annotations but also worth checking. For example, for the second claim, our model generates the question *"Did any other states have a spike in coronavirus cases related to voting?"* Although the gold fact-check did not address this question, this kind of context is the kind of thing a fact-checker may want to be attentive to, even if the answer ends up being no, and we judge this to be a reasonable question to ask given only the claim.

I More examples of QABriefs

We include more examples reflecting the annotation difference between our method and QABriefs in Figure 11.

J Datasheet for CLAIMDECOMP

J.1 Motivation for Datasheet Creation

Why was the dataset created? Despite the progress made in automating the fact-checking process, the performance achieved by current models is relatively poor. Systems in this area fundamentally need to be designed with an eye towards human verification, motivating our effort to build more explainable models so that the explanations can be used to interpret a model's behavior. Therefore, we create this dataset to facilitate future research to achieve this goal. We envision that by verifying each question, we can compose the final veracity of the claim in inherently explainable way.

Has the dataset been used already? The dataset has not been used beyond the present paper, where it was used to train a question generation model and in several evaluation conditions.

Who funded the dataset? This dataset was funded by Good Systems,[10] a UT Austin Grand Challenge to develop responsible AI technologies.

J.2 Dataset Composition

What are the instances? Each instance is a real-world political claim. All claims are written in English and most of them are US-centric.

[10] https://goodsystems.utexas.edu/

How many instances are there? Our dataset consists of two-way annotation of 1,200 claims, and 6,555 decomposed questions. A detailed breakdown of the number of instances can be seen in Table 1 of the main paper.

What data does each instance consist of? Each instance contains a real-world political claim and a set of yes-no questions with associated answers.

Does the data rely on external resources? Yes, assembling it requires access to PolitiFact.

Are there recommended data splits or evaluation measures? We include the recommended train, development, and test sets for our datasets. The distribution can be found in Table 1.

J.3 Data Collection Process

How was the data collected? We recruit 8 annotators with background in literature or politics from the freelancing platform Upwork. Given a claim paired with the justification written by the professional fact-checker on PolitiFact, we ask our annotators to reverse engineer the fact-checking process: generate yes-no questions which are answered in the justification part. For each question, the annotators also give the answer and select the relevant text in the justification that is used for the generation. The annotators are instructed to cover as many of the assertions made in the claim as possible without being overly specific in their questions.

Who was involved in the collection process and what were their roles? The 8 annotators we recruited perform the all the annotation steps outlined above.

Over what time frame was the data collected? The dataset was collected over a period from January to April 2022.

Does the dataset contain all possible instances? Our dataset does not cover all possible political claims. It mainly include complex political claims made by notable political figures of the U.S. through 2012 to 2021.

If the dataset is a sample, then what is the population? It represents a subset of all possible complex political claims which require verifying multiple aspects of the claim to reach a final veracity. Our dataset also only includes claims written in English.

J.4 Data Preprocessing

What preprocessing / cleaning was done? We remove any additional whitespace in the annotated questions, but otherwise we do not postprocess the annotations in any way.

Was the raw data saved in addition to the cleaned data? Yes

Does this dataset collection/preprocessing procedure achieve the initial motivation? Our collection process indeed achieves our initial goals of creating a high-quality dataset of complex political claims with the decompositions in question form. Using this data, we are able to check the explicit and implicit arguments made by the politicians.

J.5 Dataset Distribution

How is the dataset distributed? We make our dataset available at https://jifan-chen.github.io/ClaimDecomp.

When was it released? Our data and code is currently available.

What license (if any) is it distributed under? CLAIMDECOMP is distributed under the CC BY-SA 4.0 license.[11]

Who is supporting and maintaining the dataset? This dataset will be maintained by the authors of this paper. Updates will be posted on the dataset website.

J.6 Legal and Ethical Considerations

Were workers told what the dataset would be used for and did they consent? Crowd workers informed of the goals we sought to achieve through data collection. They also consented to have their responses used in this way through the Amazon Mechanical Turk Participation Agreement (note that even though we recruited through Upwork, workers performed annotation in the Mechanical Turk sandbox).

If it relates to people, could this dataset expose people to harm or legal action? Our dataset does not contain any personal information of crowd workers. However, our dataset can include incorrect information in the form of false claims. These claims were made in a public setting by notable political figures; in our assessment, such claims are

[11] https://creativecommons.org/licenses/by-sa/4.0/legalcode

already notable and we are not playing a significant role in spreading false claims as part of our dataset. Moreover, these claims are publicly available on PolitiFact along with expert assessment of their correctness. We believe that there is a low risk of someone being misled by information they see presented in our dataset.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? We acknowledge that, because our dataset only covers English and annotators are required to be located in the US, our dataset lacks representation of claims that are relevant in other languages and to people around the world. The claims themselves could reflect misinformation rooted in racism, sexism, and other forms of intergroup bias.

Annotation Instructions (Click to collapse)

Instructions:

Thank you for participating in this task! The goal of this task is to identify the reasoning process of a fact-checker when checking complex claims made by politicians. The current batch is a pilot, we will adjust the task and scale the task in the future according to your annotations :)

We will show you several paragraphs written by a professional fact-checker breaking down reasons why they think a claim is true or false. Your task is to identify the major questions that they are answering in this paragraph, specifically using **binary (yes/no) questions.**

Your question should ideally be one that's **motivated by the original claim**. This claim was what the fact-checkers were checking, so it was the starting point for their analysis.

The questions should not be overly specific. For example, if the analysis describes how unemployment fell by 5% over a six-year time period, the question "Did unemployment fall over this period?" is better than "Did unemployment fall by 5% over a six-year time period?" The first question is probably the one that the fact-checker set out to answer, and the specific statistics are just part of that answer to the question.

Add questions from claim if the reasoning part doesn't cover everything: Sometimes the reasoning part only checks the most important aspects of the claim leaving some minor aspects unchecked. In such cases, you should add questions according to the claim to make sure all aspects in the claim is covered by the questions.

Below we provide examples to help you better understand the task.

Claim	Decision making justification	Questions	Explanation
Barry DuVal stated on September 25, 2015 in an interview: "We're the only major oil-producing nation in the world with a self-imposed ban on exporting our crude oil to other nations."	"DuVal said the U.S. is the only major oil-producing nation in the world that bans export of its crude oil. Two experts we contacted agreed with DuVal's statement, and officials at the EIA said they're not aware of any other country with similar export restrictions. But the ban is not absolute -- a small portion of U.S. crude is exported to Canada." So we rate his DuVal's statement Mostly True.	Is the U.S. the only major oil-producing nation to ban exports of crude oil? (Yes)	The first sentence of the justification is just a restatement of the claim, so we don't write any questions. This question is mainly based on the highlighted sentence of the justification. The main point of the sentence is that the U.S. is the only country that has a ban on exporting our crude oil. This also reflects what's expressed in the original claim. We think answering this question is helpful to check the original claim.
Barry DuVal stated on September 25, 2015 in an interview: "We're the only major oil-producing nation in the world with a self-imposed ban on exporting our crude oil to other nations."	"DuVal said the U.S. is the only major oil-producing nation in the world that bans export of its crude oil. Two experts we contacted agreed with DuVal's statement, and officials at the EIA said they're not aware of any other country with similar export restrictions. But the ban is not absolute -- a small portion of U.S. crude is exported to Canada." So we rate his DuVal's statement Mostly True.	Is the U.S. ban on crude oil export a complete ban? (No)	Based on the highlighted part, we see it adds extra information over the crude oil ban. Although it is not explicitly mentioned in the original claim, the fact-checker felt that answering this question was important to give more context to the claim. The question "Is a small portion of US crude exported to Canada?" is not as good. **Since Canada is not presented in the original claim**, this was probably not what the fact-checker set out to answer; they only discovered it after doing their research.
A Facebook post stated on January 31, 2021: "Nancy Pelosi bought $1.25 million in Tesla stock the day before Joe Biden signed an order "for all federal vehicles" to be electric."	"An image shared on Facebook claims that Nancy Pelosi bought $1.25 million in Tesla stock the day before Biden signed an order "for all federal vehicles" to be electric, implying that she sought to profit from inside information about new government policies. The House speaker did report transactions involving Tesla stock, but the post misrepresented the purchases and Biden's policies to create the false impression that the transactions represented improper insider trading in Tesla shares. The statement contains an element of truth, but ignoring critical facts would give a different impression.	Were the stock purchases improper insider trading? (No)	The first part of this sentence talks about that the the stock purchases and Biden's policy were misrepresented, but both the purchase and the policy are mentioned in the original claim. Therefore, we don't ask the questions about the two parts here. This sentence also mentions the claim gives a false impression that this purchase involves insider trading, so we ask the above question here.
A Facebook post stated on January 31, 2021: "Nancy Pelosi bought $1.25 million in Tesla stock the day before Joe Biden signed an order "for all federal vehicles" to be electric."	"An image shared on Facebook claims that Nancy Pelosi bought $1.25 million in Tesla stock the day before Biden signed an order "for all federal vehicles" to be electric, implying that she sought to profit from inside information about new government policies. The House speaker did report transactions involving Tesla stock, but the post misrepresented the purchases and Biden's policies to create the false impression that the transactions represented improper insider trading in Tesla shares. The statement contains an element of truth, but ignoring critical facts would give a different impression.	Does the executive order Biden signed require all federal vehicles to be electric? (unknown)	Beyond the stock purchases, we need to check whether there actually was an order from Biden about electric vehicles. As the answer is not obvious from the reasoning part, we give "unknown" here. We feel like these three questions covered the reasoning that the fact-checker wrote. It seems like these were the two most salient aspects they addressed.

Claim

Barry DuVal stated on September 25, 2015 in an interview: " We 're the only major oil - producing nation in the world with a self - imposed ban on exporting our crude oil to other nations . "

Justification

"DuVal said the U.S. is the only major oil-producing nation in the world that bans export of its crude oil.

Two experts we contacted agreed with DuVal's statement, and officials at the EIA said they're not aware of any other country with similar export restrictions. But the ban is not absolute -- a small portion of U.S. crude is exported to Canada."

According to the claim and its justification above, write down one or more binary questions (answerable by yes/no) that is answerable by the justification part. Remember the three rules: (1) Write questions that are **helpful to verify the original claim**. (2) **Don't write questions that are too specific**. You should also provide an answer to your question -- yes/no. (3) Add questions from claim if the reasoning part doesn't cover everything. Also, you should copy-paste the justification text you used to generate the question (usually one sentence).

Many claims have around **3 questions** that are being addressed. We are most interested in collecting a comprehensive set of these questions, so we encourage you to write a few questions. Use the following buttons to add/remove inputs.

+ -

question:

write down your question

answer:

yes, no, unknown

text used to generate the question:

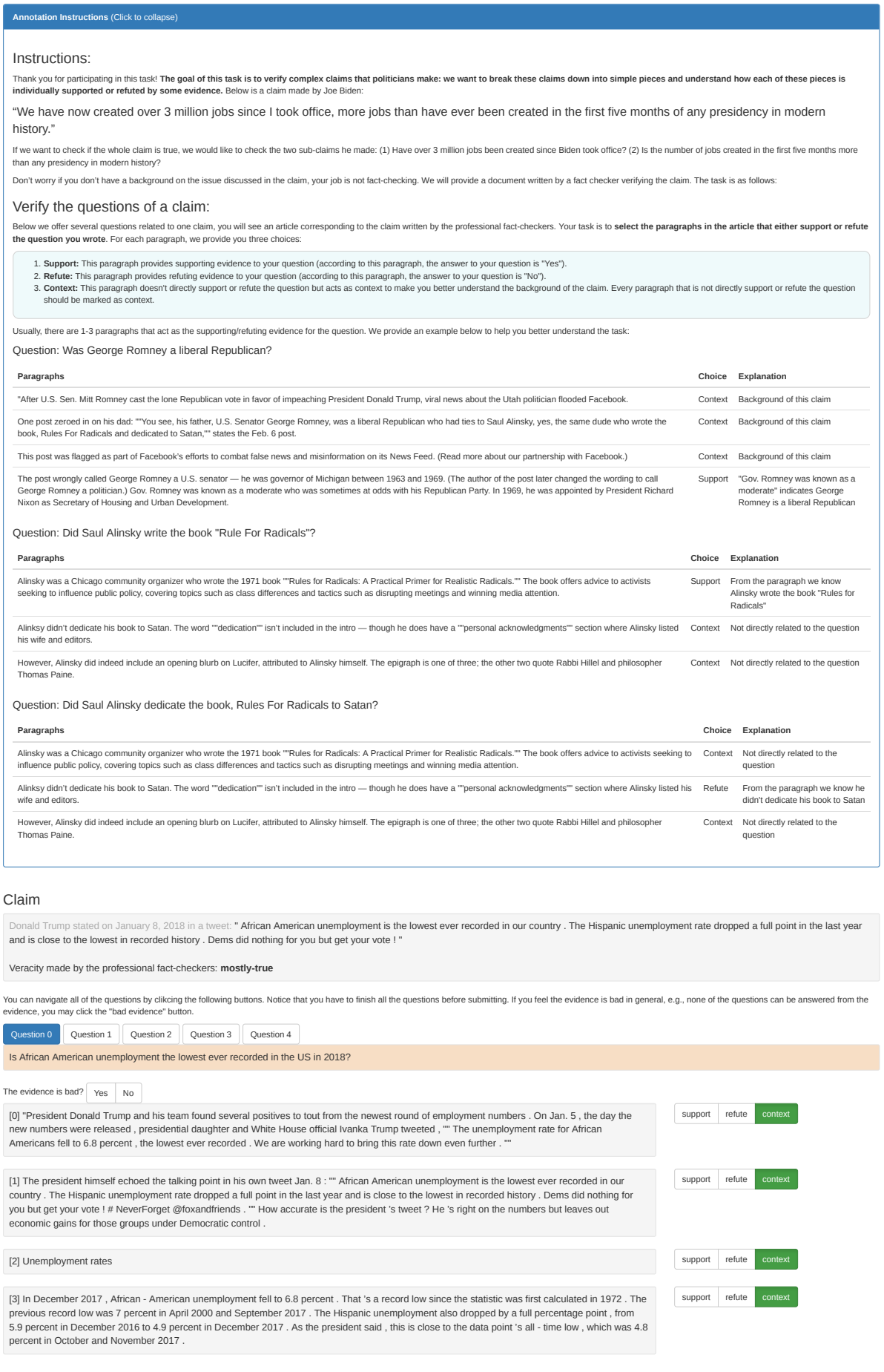
copy-paste the related text

text from reasoning or claim:

reasoning/claim

Submit

Figure 7: Interface of our question decomposition task, including the annotation instructions.



Annotation Instructions (Click to collapse)

Instructions:

Thank you for participating in this task! **The goal of this task is to verify complex claims that politicians make: we want to break these claims down into simple pieces and understand how each of these pieces is individually supported or refuted by some evidence.** Below is a claim made by Joe Biden:

"We have now created over 3 million jobs since I took office, more jobs than have ever been created in the first five months of any presidency in modern history."

If we want to check if the whole claim is true, we would like to check the two sub-claims he made: (1) Have over 3 million jobs been created since Biden took office? (2) Is the number of jobs created in the first five months more than any presidency in modern history?

Don't worry if you don't have a background on the issue discussed in the claim, your job is not fact-checking. We will provide a document written by a fact checker verifying the claim. The task is as follows:

Verify the questions of a claim:

Below we offer several questions related to one claim, you will see an article corresponding to the claim written by the professional fact-checkers. Your task is to **select the paragraphs in the article that either support or refute the question you wrote**. For each paragraph, we provide you three choices:

1. **Support:** This paragraph provides supporting evidence to your question (according to this paragraph, the answer to your question is "Yes").
2. **Refute:** This paragraph provides refuting evidence to your question (according to this paragraph, the answer to your question is "No").
3. **Context:** This paragraph doesn't directly support or refute the question but acts as context to make you better understand the background of the claim. Every paragraph that is not directly support or refute the question should be marked as context.

Usually, there are 1-3 paragraphs that act as the supporting/refuting evidence for the question. We provide an example below to help you better understand the task:

Question: Was George Romney a liberal Republican?

Paragraphs	Choice	Explanation
"After U.S. Sen. Mitt Romney cast the lone Republican vote in favor of impeaching President Donald Trump, viral news about the Utah politician flooded Facebook.	Context	Background of this claim
One post zeroed in on his dad: ""You see, his father, U.S. Senator George Romney, was a liberal Republican who had ties to Saul Alinsky, yes, the same dude who wrote the book, Rules For Radicals and dedicated to Satan,"" states the Feb. 6 post.	Context	Background of this claim
This post was flagged as part of Facebook's efforts to combat false news and misinformation on its News Feed. (Read more about our partnership with Facebook.)	Context	Background of this claim
The post wrongly called George Romney a U.S. senator — he was governor of Michigan between 1963 and 1969. (The author of the post later changed the wording to call George Romney a politician.) Gov. Romney was known as a moderate who was sometimes at odds with his Republican Party. In 1969, he was appointed by President Richard Nixon as Secretary of Housing and Urban Development.	Support	"Gov. Romney was known as a moderate" indicates George Romney is a liberal Republican

Question: Did Saul Alinsky write the book "Rule For Radicals"?

Paragraphs	Choice	Explanation
Alinsky was a Chicago community organizer who wrote the 1971 book ""Rules for Radicals: A Practical Primer for Realistic Radicals."" The book offers advice to activists seeking to influence public policy, covering topics such as class differences and tactics such as disrupting meetings and winning media attention.	Support	From the paragraph we know Alinsky wrote the book "Rules for Radicals"
Alinksy didn't dedicate his book to Satan. The word ""dedication"" isn't included in the intro — though he does have a ""personal acknowledgments"" section where Alinsky listed his wife and editors.	Context	Not directly related to the question
However, Alinsky did indeed include an opening blurb on Lucifer, attributed to Alinsky himself. The epigraph is one of three; the other two quote Rabbi Hillel and philosopher Thomas Paine.	Context	Not directly related to the question

Question: Did Saul Alinsky dedicate the book, Rules For Radicals to Satan?

Paragraphs	Choice	Explanation
Alinsky was a Chicago community organizer who wrote the 1971 book ""Rules for Radicals: A Practical Primer for Realistic Radicals."" The book offers advice to activists seeking to influence public policy, covering topics such as class differences and tactics such as disrupting meetings and winning media attention.	Context	Not directly related to the question
Alinksy didn't dedicate his book to Satan. The word ""dedication"" isn't included in the intro — though he does have a ""personal acknowledgments"" section where Alinsky listed his wife and editors.	Refute	From the paragraph we know he didn't dedicate his book to Satan
However, Alinsky did indeed include an opening blurb on Lucifer, attributed to Alinsky himself. The epigraph is one of three; the other two quote Rabbi Hillel and philosopher Thomas Paine.	Context	Not directly related to the question

Claim

Donald Trump stated on January 8, 2018 in a tweet: " African American unemployment is the lowest ever recorded in our country . The Hispanic unemployment rate dropped a full point in the last year and is close to the lowest in recorded history . Dems did nothing for you but get your vote ! "

Veracity made by the professional fact-checkers: **mostly-true**

You can navigate all of the questions by clikcing the following buttons. Notice that you have to finish all the questions before submitting. If you feel the evidence is bad in general, e.g., none of the questions can be answered from the evidence, you may click the "bad evidence" button.

Question 0 | Question 1 | Question 2 | Question 3 | Question 4

Is African American unemployment the lowest ever recorded in the US in 2018?

The evidence is bad? Yes | No

[0] "President Donald Trump and his team found several positives to tout from the newest round of employment numbers . On Jan. 5 , the day the new numbers were released , presidential daughter and White House official Ivanka Trump tweeted , "" The unemployment rate for African Americans fell to 6.8 percent , the lowest ever recorded . We are working hard to bring this rate down even further . ""

support | refute | context

[1] The president himself echoed the talking point in his own tweet Jan. 8 : "" African American unemployment is the lowest ever recorded in our country . The Hispanic unemployment rate dropped a full point in the last year and is close to the lowest in recorded history . Dems did nothing for you but get your vote ! # NeverForget @foxandfriends . "" How accurate is the president 's tweet ? He 's right on the numbers but leaves out economic gains for those groups under Democratic control .

support | refute | context

[2] Unemployment rates

support | refute | context

[3] In December 2017 , African - American unemployment fell to 6.8 percent . That 's a record low since the statistic was first calculated in 1972 . The previous record low was 7 percent in April 2000 and September 2017 . The Hispanic unemployment also dropped by a full percentage point , from 5.9 percent in December 2016 to 4.9 percent in December 2017 . As the president said , this is close to the data point 's all - time low , which was 4.8 percent in October and November 2017 .

support | refute | context

Figure 8: Interface of our evidence annotation task used in section 6, including the annotation instructions.

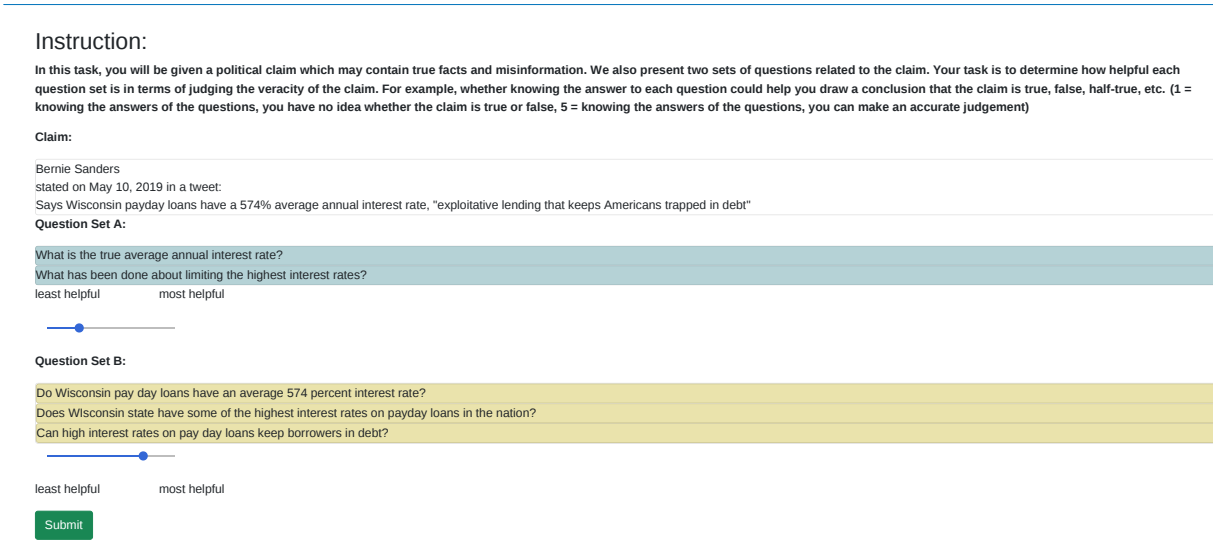


Figure 9: Interface of our user study conducted in section 5.2, including the annotation instructions.

Claim: JD Wooten stated on August 24, 2018: " My opponent , Rick Gunn , blocked the expansion of Medicaid — costing half a million people health insurance , including at least 23,000 veterans . "

Annotator A	Annotator B
Q1 : Did Rick Gunn block the expansion of Medicaid ? ☐	Q1 : Did Rick Gunn block the expansion of Medicaid ? ☐
Q2 : Did half a million people lose health insurance under the bill Rick Gunn voted for ? ☐	Q2 : Is this blockage depriving half a million people of health insurance ? ☐
Q3 : Did a sizable number of people lose health insurance under the bill Gunn voted for ? ■	Q3 : Are roughly 23,000 of these people deprived of health insurance veterans ? ☐
Q4 : Did a large number of veterans lose health insurance because of the bill Gunn voted for ? ☐	

Claim: Hillary Clinton stated on February 4, 2016 in a debate: " We bailed out " Johnson Controls when " we saved the auto industry " and " now they want to avoid paying taxes . "

Annotator A	Annotator B
Q1 : Did we directly bail out Johnson Controls when we saved the auto industry ? ☐	Q1 : Did Johnson Controls directly benefit from the auto industry bailouts ? ☐
Q2 : Did Johnson Controls benefit in any aspect from federal bailouts ? ☐	Q2 : Did Johnson Controls benefit in any way from the auto industry bailouts ? ☐
Q3 : Does Johnson Controls want to avoid paying taxes ? ☐	Q3 : Is Johnson Controls trying to avoid paying U.S. taxes ? ☐
Q4 : Did Johnson want to avoid paying taxes because of being bailed out ? ■	

Figure 10: Two examples of our inter-annotator agreement assessment: giving two set of questions, we mark questions of which the semantics cannot be matched to the other annotator's set. Here, the black box denotes the marked question.

Claim	QA-briefs (Fan et al., 2020)	CLAIMDECOMP (Ours)
Bernie Sanders stated on March 6, 2016 in a Democratic presidential debate: “Almost every poll has shown that Sanders vs. Trump does a lot better than Clinton vs. Trump and, that’s true nationally."	Why is it true that Sanders vs. Trump did a lot better than Clinton vs. Trump? Which polls have shown that Sanders vs. Trump did better than Hillary vs. Trump?	Did almost every poll since Jan 1 show that Sanders vs. Trump did a lot better than Clinton vs. Trump? Are the results of those polls accurate predictors of the November results?
Henry Cuellar stated on February 22, 2019 in a press conference: President Donald Trump "put out a $297 million contract to a vendor to show him how to hire Border Patrol. So, the first payment they made was $14.8 million. You know how many Border Patrol they hired with $14.8 million? Two."	Why would trump put out a 297 million contract to hire Patrol? Why is it so costly to hire Border Patrol?	Did a hiring vendor receive a $297 million contract to hire Border Patrol? Did this hiring vendor receive a first payment of $14.8 million? Did this hiring vendor recruit two agents with this first round of funding? Did this hiring vendor assist in recruiting other agents?
Taylor Swift stated on October 7, 2018 in an Instagram post: Marsha Blackburn "voted against the Reauthorization of the Violence Against Women Act, which attempts to protect women from domestic violence, stalking, and date rape."	Why did Marsha vote against the Re-authorization of the Violence Against Women Act? Why is Blackburn terrified by Taylor swift?	Did Marsha Blackburn vote against the Reauthorization of the Violence Against Women Act? Did the act attempt to protect women from domestic violence, stalking, and date rape? Did Blackburn vote for an alternative version of this reauthorization?

Figure 11: More examples of the comparison between our decomposed questions with QABriefs (Fan et al., 2020).

Claim	Predicted Questions	Annotated Questions
Chip Roy stated on May 6, 2020 in an interview: Wisconsin has not had a spike in coronavirus cases that was statistically significant related to the fact that they had voting.	Is there any proof Wisconsin had a spike in coronavirus cases related to voting? Did any other states have a spike in coronavirus cases related to voting? Were the spikes in coronavirus cases in Wisconsin statistically significant?	Did Wisconsin have a spike in coronavirus cases? Did Wisconsin have a spike in coronavirus cases that was related to voting? Were any coronavirus cases in Wisconsin linked to the election?
Bill Maher stated on February 10, 2016 in an essay in The Hollywood Reporter: Why are Muslim armies so useless against ISIS? ... There are 20,000 or 30,000 (ISIS fighters). The countries surrounding ISIS have armies totaling 5 million people.	Are there only 20,000 to 30,000 ISIS fighters in total? Do the countries surrounding ISIS have armies totaling 5 million people? Are the armies of the countries surrounding ISIS ineffective against ISIS? Was the number of ISIS fighters in total greater than the armies of the countries surrounding ISIS?	Do countries sharing a direct border with ISIS strongholds have 5 million troops? Are there 20,000 or 30,000 (ISIS fighters)? Does the figure for total armies include reserve troops? Are these countries surrounding ISIS politically involved with them?
Charlie Kirk stated on February 24, 2020 in a tweet: “Three Chinese nationals were apprehended trying to cross our Southern border illegally. Each had flu-like symptoms. Border Patrol quickly quarantined them and assessed any threat of coronavirus.”	Were these Chinese nationals quarantinated once they were apprehended? Were these Chinese nationals suspected of having the coronavirus? Have any other coronavirus cases been reported in this area? Would this group of people be considered a high risk for the coronavirus?	Did three people from China attempt to cross the southern border illegally? Did all three people from China have flu-like symptoms? Did Border Patrol quarantine the three people from China? Did any of the three Chinese people have the coronavirus?

Figure 12: Questions generated by the QG-MULTIPLE model, compared with the annotations.