Complex Claim Verification with Evidence Retrieved in the Wild

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

Abstract

Retrieving evidence to support or refute claims is a core part of automatic fact-checking. Prior work makes simplifying assumptions in retrieval that depart from real-world use cases: either no access to evidence, access to evidence curated by a human fact-checker, or access to evidence published after a claim was made. In this work, we present the first realistic pipeline to check real-world claims by retrieving raw evidence from the web. We restrict our retriever to only search documents available prior to the claim's making, modeling the realistic scenario of emerging claims. Our pipeline includes five components: claim decomposition, raw document retrieval, fine-grained evidence retrieval, claim-focused summarization, and veracity judgment. We conduct experiments on complex political claims in the CLAIMDECOMP dataset and show that the aggregated evidence produced by our pipeline improves veracity judgments. Human evaluation finds the evidence summary produced by our system is reliable (it does not hallucinate information) and relevant to answering key questions about a claim, suggesting that it can assist fact-checkers even when it does not reflect a complete evidence set.[1]

1 Introduction

To combat the rise of misinformation, the NLP community has developed automatic fact-checking tools. However, these automated systems are not ready for wide adoption at real fact-checking organizations. Prior work handling real claims either relies on access to a document set which contains the "gold" evidence (Ferreira and Vlachos, 2016; Alhindi et al., 2018; Hanselowski et al., 2019; Atanasova et al., 2020) or conducts unconstrained retrieval (Augenstein et al., 2019), which may retrieve articles written by fact-checkers about the claim (example in Figure 1).

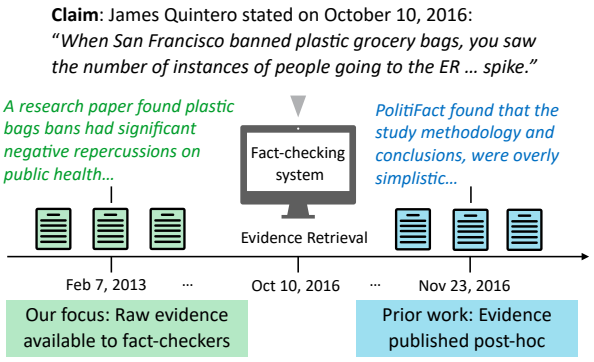


Figure 1: Our fact-checking setting addresses realistic claims using evidence retrieved prior to when the claim was made.

We present the first study of fact-checking political claims under a realistic retrieval setting. Our retrieval over the web is restricted to documents authored before the time of the claim and not sourced from fact-checking websites, as shown by the left side of Figure 1. We propose a pipeline (illustrated in Figure 2) that builds upon prior work in fact checking as well as large language models (Brown et al., 2020) to handle the complexity of this setting. Our system first decomposes a claim into a series of subquestions (Chen et al., 2022a; Ousidhoum et al., 2022), targeting both explicit and implicit aspects of the claim. Each subquestion is fed into a commercial search engine to retrieve relevant documents, with the restrictions described above. Then, we conduct a second stage of fine-grained retrieval to isolate the most relevant portions of the documents. Finally, we use state-of-the-art language models (Brown et al., 2020; Ouyang et al., 2022) to generate claim-focused summaries from the retrieved content. These summaries can serve both as explanations for users as well as inputs to a classifier to determine the veracity based on these summaries.

Evaluating individual components of our

[1]Code and data available at https://github.com/jifan-chen/Fact-checking-via-Raw-Evidence

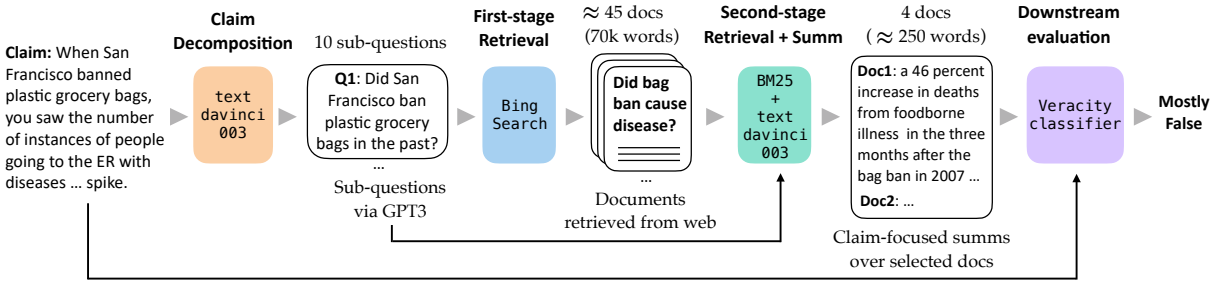


Figure 2: Overview of our pipeline: a claim is decomposed into yes/no subquestions (Sec. 3.1), then we use the questions in two stages of retrieval (Sec. 3.2 and Sec. 3.3) to select the most relevant paragraphs. Finally, we generate a claim-focused summary (Sec. 3.4) and train a veracity classifier to get the veracity label (Sec. 3.5). This filters contents irrelevant to the claim (see Appendix B for details and an example in Figure 5).

pipeline is challenging due to the absence of gold annotations at each stage. We use automatic evaluation on the veracity classification performance, comparing to labels given by professional fact-checkers. We supplement this with a human study evaluating the claim-focused summaries for comprehensiveness and faithfulness. This evaluation counterbalances the subjectivity of the veracity judgments (Lim, 2018) while shedding light on intermediate stages of the process.

We apply our pipeline to CLAIMDECOMP (Chen et al., 2022a), a dataset containing 1,200 real-world complex political claims with veracity labels. Performance on veracity classification shows that: (1) our retrieval setting is indeed much harder than "unrestricted" retrieval settings; (2) using web evidence leads to performance gains compared to automatic fact-checking without evidence; (3) the decomposition is crucial for obtaining high-quality raw documents from the web compared to using the original claim alone. Our human study further indicates that: (4) claim-focused summaries are mostly faithful and helpful for both machines and humans to fact-check a claim; (5) the retrieved evidence is often relevant to some aspects of the claim, but can rarely cover all aspects, suggesting that finding sufficient raw evidence in the wild is the core challenge in building automatic fact-checking systems.

2 Background and Motivation

Early NLP research on fact-checking political claims (Vlachos and Riedel, 2014; Wang, 2017; Rashkin et al., 2017; Volkova et al., 2017; Pérez-Rosas et al., 2018; Dungs et al., 2018) typically considered using the claim alone as an input to an automated system. By not seeking evidence, systems judge the veracity of a claim mostly based on surface-level linguistic patterns rather than based on factual errors. Research that incorporates evidence either assumes access to justifications provided by fact-checkers (Vlachos and Riedel, 2014; Alhindi et al., 2018; Hanselowski et al., 2019; Atanasova et al., 2020) or evidence from *unconstrained* retrieval (Popat et al., 2017, 2018; Augenstein et al., 2019), which frequently yields evidence sets containing pages from fact-checking websites (Glockner et al., 2022). This does not reflect the difficulties in real-world evidence retrieval. Fan et al. (2020) explore generating questions to retrieve evidence from the web, but only evaluate their system with humans in the loop, who can aggressively filter irrelevant retrieval results. Contemporaneous to this work, Schlichtkrull et al. (2023) construct a dataset, AVeriTeC, using real-world claims and evidence retrieved from the web. Our method uses binary subquestions designed to target all needed aspects of factuality for a claim, whereas their questions are wh-questions optimized around retrieval, similar to QABriefs (Fan et al., 2020).

To our knowledge, we present the first automatic fact-checking system with a realistic retrieval pipeline using evidence available at the time a claim was made. This presents a very challenging setting where many claims are not checkable. We therefore emphasize the evidence our system returns as a way of assisting human fact-checkers; we believe this realistic task setting and corresponding evaluation should be reused in future work.

Our work shifts the focus away from the evaluation on classification accuracy alone. Accuracy on truth labels assigned by fact-checkers is a proxy metric we use to evaluate our systems. However, fact-checking experts argue that the task is too subjective and complex to be automated in the near term (Graves, 2018; Nakov et al., 2021). Part of this arises from the fact that information needed to check claims is not always available on the web (Singh et al., 2021). Our approach of returning

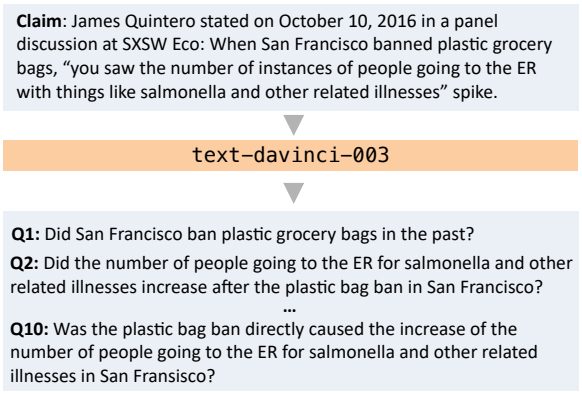


Figure 3: An example of our claim decomposition process: each claim is decomposed into ten subquestions.

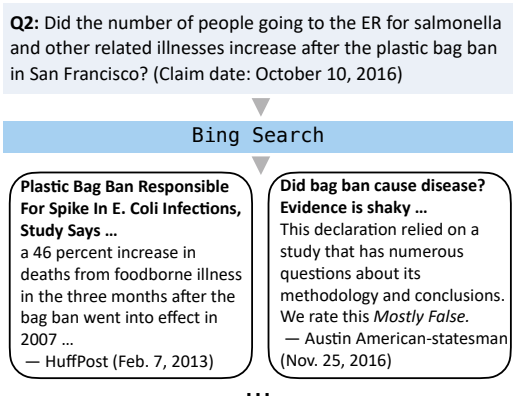


Figure 4: Two documents returned by searching Q2 (generated in step 1). The right page post-dates the claim by one month and directly cites a PolitiFact article, making it problematic to use as raw evidence.

information on a best-effort basis and providing evidence to enable humans to assist in the judgment can help overcome issues with returning judgments from error-prone AI systems (Bansal et al., 2021; Brand et al., 2022).

3 Methodology

Our pipeline, shown in Figure 3, consists of five parts: claim decomposition, raw document retrieval, fine-grained retrieval, claim-focused summarization, and veracity classification. We describe each part below.

3.1 Claim Decomposition

Given a real-world complex claim, we first decompose it into a set of yes/no questions for which the answers are useful to fact-check the claim. Chen et al. (2022a); Ousidhoum et al. (2022) show that such decompositions are both helpful to retrieve relevant evidence and make veracity judgments.

For decomposition, we prompt a large-scale language model, `text-davinci-003`, with in-context examples.[2] We carefully choose four input-decomposition pairs from the human annotations of Chen et al. (2022a) to form a few-shot prompt. We generate a set of questions through multiple rounds of sampling until we gather 10 different questions. An example decomposition is shown in Figure 3. For the full prompt, see Appendix A.2.

[2]During a pilot study, we compared the questions generated from `text-davinci-003` and the questions generated using the fine-tuned T5-3b model from Chen et al. (2022a) and we found that the questions generated by `text-davinci-003` are more diverse and comprehensive.

3.2 First-stage Retrieval

For each question generated in the previous step, we feed it to a commercial search engine API to collect the relevant documents.

Temporal and Site Constraints We assume that **a system should not be able to access pages published after the claim was made.** This condition matches real-time fact-checking scenario during a political speech. We place a **temporal constraint** on the system to reflect this. Next, to investigate how the presence of fact-checking websites affects the veracity judgment of a claim, we also place a **site constraint** to filter out the documents from fact-checking websites. Our list of fact-checking websites can be found in Appendix A.1. An example of the retrieved documents is shown in Figure 4.

We use the Bing Search API,[3] and retrieve 10 documents per subquestion after filtering by the constraints. We extract the actual content from the page URLs using two tools: `html2text`[4] and `readability-lxml`.[5] Approximately one-third of the URLs are protected[6] and cannot be scraped.

Table 1 contains the raw counts from web retrieval with and without the timestamp of a claim. These results underscore the importance of temporal filtering: we find little overlap between the two document sets by comparing the Jaccard distance between two sets of the retrieved URLs.

One challenge for the reproducibility of our work

[3]http://www.microsoft.com/en-us/bing/apis/bing-web-search-api

[4]https://github.com/Alir3z4/html2text/

[5]https://github.com/buriy/python-readability

[6]Paywall, PDFs, and anti-scraping measures.

Claim: Melissa Agard stated on September 2, 2021 in News release: "No other country on the planet witnesses the number of gun deaths that we do here in the United States, and it's not even close."

Decomposed subquestions:
(1) Is the United States the country with the highest rate of gun deaths?
(2) Does the claim account for population size (i.e., per capita rates), or is it based on total numbers?
(3) Does the statement consider gun deaths relative to the total number of guns in the country?
(4) Is the number of gun deaths in the United States substantially higher when compared to countries of similar economic and political stability?
(5) Do gun deaths account for a large portion of deaths in the U.S.?
...
(10) Are there any mitigating factors that affect the gun death rate in the United States?

Retrieved documents and summaries:

Doc title	URL	Summary
Firearm-related deaths rate U.S. by gender 1970-2016 | Statista (Nov 7, 2019)	https://www.statista.com/statistics/186951/deaths-by-firearm-related-injuries-in-the-us-by-gender-since-1970/	The death rate from firearm-related injuries in the United States is 19.4 per 100,000 population among males, and homicides from firearms account for 72.6 percent of all homicides in the U.S. The ownership of legal firearms is widespread, with around 43 percent of households having at least one firearm. (Faithful)
Do Gun Laws Affect the Rate of Shooting Deaths? (Oct 12, 2018)	https://www.thetrace.org/2018/10/do-gun-laws-affect-the-rate-of-shooting-deaths/	The Centers for Disease Control and Prevention tracks gun deaths in all 50 states and the national rate of gun deaths in 2016 was 11.8. **No additional evidence is provided to compare the number of gun deaths in the US to other countries.** (Faithful)
Gun Violence Deaths: How The U.S. Compares With The Rest Of The World (Mar 24, 2021)	https://www.kuow.org/stories/gun-violence-deaths-how-the-u-s-compares-with-the-rest-of-the-world	**The United States has the 32nd highest rate of deaths from gun violence with 3.96 deaths per 100,000 people** compared to countries such as Canada with 0.47 deaths per 100,000 people and the United Kingdom with 0.04 deaths per 100,000 people. (Faithful)
A Doctor's Insights Into Gun Violence And Gun Laws Around The World (Aug 6, 2019)	https://www.kuer.org/2019-08-06/a-doctors-insights-into-gun-violence-and-gun-laws-around-the-world	The US rate of deaths from gun violence is **4.43 deaths per 100,000 people and is four times higher than the rates in war-torn Syria and Yemen**. This places is among the top 30 countries in the world with the highest rates of deaths from gun violence. (Faithful)

Annotated subquestions:
Does the U.S. have the highest number of gun deaths out of all the countries on the planet? **No** (annotator judgment based on summaries)
Does the U.S. have a high number of gun deaths? **Yes** (annotator judgment based on summaries)
Does the U.S. have a high number of gun deaths when looking at deaths as a share of the population? **Yes** (annotator judgment based on summaries)

Model Prediction: Half-True

Label: Mostly-False

Figure 5: System outputs for an example picked from the dev set of CLAIMDECOMP: the claim is first decomposed into a set of yes/no questions and then the top four retrieved documents (through first and second stage retrieval) are summarized. Finally, a trained DeBERTa model makes a prediction regarding the four summarized documents.

	# retrieved	# scraped	# words
w/ timestamp	66.7	45.0	1,561
w/o timestamp	70.4	47.8	1,660
Jaccard score	0.12	0.12	-

Table 1: The statistics for the retrieved documents obtained through the first-stage retrieval after filtering the documents from fact-checking websites. Jaccard between these two sets show that incorporating the timestamp in retrieval makes a substantial difference.

is that commercial search engines may return different results over time. In Section 5.3, we experiment with the same query set at different times. We find that the search results change over time: only 30% of search output URLs overlap when queried two months apart. However, the veracity judgment classification result is not impacted much.

3.3 Second-stage Retrieval

Most of the documents collected from the previous step contain at most a few snippets relevant to the claim. However, as can be seen from Figure 2, first-stage retrieval can easily result in tens of thousands of words of retrieved documents, which are costly to process with an LLM. Furthermore, even with state-of-the-art language models, it is hard to do complex reasoning over such long context (Liu et al., 2024; Levy et al., 2024). Thus, we conduct a second-stage retrieval to pick the most relevant text spans to the claim from the retrieved documents. Specifically, we segment the documents into text spans containing k_1 words with a stride of $\frac{1}{2}k_1$ words. Following Chen et al. (2022a), we employ `BM-25` to retrieve the top-K_1 highest-scored text spans, expanding these spans with a $\pm$ k_2-word context. If two text spans overlap, they are merged to form a larger span. This process yields a set of "documents" ranked by the highest-scored text spans, of which we pick the top-K_2.

3.4 Claim-Focused Summarization

Since the documents retrieved in the previous step can contain up to several thousand words, it be-

comes cumbersome for both humans and models to make a judgment based on them (Stammbach and Ash, 2020). Consequently, we prompt a large language model, specifically `text-davinci-003`, to summarize each retrieved document *separately* with respect to the claim.[7] Such single-document summarization has been shown to be robust on news articles (Goyal et al., 2022; Zhang et al., 2023).

We investigate two types of prompts. For a **zero-shot** prompt, we instruct the model not to make any judgments about the stance of the given document. For a **few-shot** prompt, we select four documents and carefully write desired summaries. For documents that are not relevant to the claim, we write "the document is not relevant to checking the claim" as its desired output. We conduct human evaluation of the summary quality of different prompts in Section 6.1, where we find that few-shot prompting works better. See Appendix A.3 for full prompts.

3.5 Veracity Classification

The final stage of our pipeline involves making a judgment based on the summaries generated in the previous stage. Unlike previous stages which use off-the-shelf tools, here we *train* a DeBERTa-large (He et al., 2020) model[8] to perform a six-way veracity classification (true, mostly true, half true, barely true, false, and pants-on-fire).

Training We run our pipeline over the training, development, and test data of CLAIMDECOMP and train on pairs of the form (claim+summary, label). Since the dataset is small, we train the classifier five times with different random seeds and report the test set performance using the model that achieves the best performance on the development set.

3.6 Final Pipeline

Our complete pipeline's results when executed on an example are shown in Figure 5. We note that the question decomposition phase yields an overcomplete set of questions, including redundant ones. However, the final retrieved and summarized documents are able to shed light on the claim from several complementary perspectives. While the final veracity judgment does not exactly match the judgment from PolitiFact, reading the documents still gives an informed picture of the situation.

[7]During a pilot study, we explored prompting `text-davinci-003` to generate one summary using all documents. However, it frequently went beyond simple summarization and produced verdicts such as "therefore, the claim is refuted by the document," which were unreliable compared to using our veracity classifier.

[8]We also experimented with using ChatGPT as the veracity classifier. We describe results and analysis in Appendix D; we found it yielded worse performance than the fine-tuned model.

4 Experimental Setup

Our main automatic evaluation is on claim veracity prediction (Wang, 2017), evaluating our entire pipeline end-to-end. We will describe the human evaluation setup in Section 6.

Data We use the data from CLAIMDECOMP (Chen et al., 2022a) which contains 1,200 complex claims from PolitiFact (train: 800, dev: 200, test: 200). Each claim is labeled with one of the six veracity labels, a justification paragraph written by expert fact-checkers, and subquestions annotated by prior work.

Hyperparameters For the second-stage retrieval, we set top-$K_1 = 10$ (highest-scored text spans), top-$K_2 = 4$ (highest-scored documents), $k_1 = 30$ (chunk size), and $k_2 = 150$ (expansion parameter). See appendix A.4 for all hyperparameters.

Evaluation Metric We report accuracy (Acc), mean absolute error (MAE, on our 6-point scale), and Macro-F1. We also introduce soft accuracy (soft Acc), which is calculated by counting off-by-one errors on the six-point veracity scale (e.g., *half true* instead of *mostly true*) as correct, as veracity judgments are subjective.

Comparison Systems For our **Claim-only** system, we concatenate the metadata, including the speaker and the venue of the claim, with the claim itself, and feed the resulting text into the classifier (Wang, 2017). This approach serves as a lower bound for the veracity classification.

We extend the Claim-only baseline to **Claim+Justification** by appending the human-written justification paragraph, excluding the sentence containing the label, to the claim. This is an oracle setting to establish an upper bound for veracity classification.

5 Automatic Evaluation: Claim Veracity

5.1 Constrained vs. Unconstrained Search

We first situate our work with respect to baselines and past systems by varying the retrieval condition. We experiment with a **temporal** constraint, where pages must originate before the date of the claim, and a **site** constraint, where sites must be

Retrieval Constraint		Dev (N=200)				Test (N=200)			
Temporal	Site	Acc	Soft Acc	Macro-F1	MAE	Acc	Soft Acc	Macro-F1	MAE
-	-	50.5	88.5	47.5	0.62	49.0^{+}	86.0^{+}	48.5^{+}	0.68^{+}
-	Non-FC	37.5	76.5	38.6	0.94	33.5^{+}	75.0^{+}	33.9^{+}	0.95^{+}
Before	-	42.5	75.0	41.7	0.87	33.5^{+}	72.0	38.0^{+}	0.98^{+}
Before	Non-FC	40.5	76.5	41.4	0.87	33.0^{+}	74.5^{+}	34.5^{+}	0.99^{+}
Claim only		37.0	71.0	34.6	0.98	25.5	68.0	27.5	1.12
Claim + Justification (oracle)		52.5	88.5	54.5	0.64	57.5	93.0	57.8	0.50

Table 2: Veracity classification performance with different retrieval constraints. The top block is our full system (Ⓑ setting in Table 3) with constraints over what is retrieved. Red indicates using oracle information. "+" denotes that the results are statistically significant improvements ($p < 0.05$) compared to the results of Claim only on the test set.

non-fact-checking (non-FC) sites. Even in the unconstrained setting, we exclude pages from PolitiFact (our dataset's source) to prevent label leakage.

The unconstrained setting corresponds to that used in MultiFC (Augenstein et al., 2019). **MultiFC includes numerous documents that are filtered out by our constrained settings.** For each claim, they extract the top 10 pages from the Google search API. We find that 12,721 out of 15,379 claims (82.7%) contain at least one page from our excluded website list and 24.4% of the retrieved web pages are from fact-checking websites.

Table 2 reports the performance of our system with various retrieval constraints. Comparing the performance of *claim-only* and other models that use retrieval, we see a statistically significant[9] improvement over all four of our metrics in nearly all settings, showing that **retrieving and summarizing evidence is helpful to predict the veracity label, even with constraints**.

Second, we see **adding either temporal or site constraints dramatically reduces the performance**. This implies that retrieval over the web works largely because it retrieves fact-checks that were published after the claim was released, with synthesized evidence. We believe that future work on retrieval should use a constrained setting.

5.2 Stage Ablations

We evaluate design choices in each stage of the pipeline to understand how each individual component contributes to the final performance. The results are shown in Table 3.

First-stage Retrieval: subquestions vs. original claim Using the original claim instead of the generated subquestions as an input to web search (Ⓑ vs. ①) results in a notable decrease in performance. The subquestion set encompasses multiple aspects of the claim, enabling the search engine to locate relevant information more easily across separate search queries. Comparing Ⓑ and ②, we see using the gold subquestions actually yields worse performance than our predicted subquestions. This could be because we predict 10 subquestions, potentially garnering more relevant data than the 3 (on average) gold subquestions (Chen et al., 2022a).

Second-stage Retrieval Rather than retrieving with subquestions (**subQs**), we instead perform our search with the raw **Claim** (③), **Gold subQs** from CLAIMDECOMP (④), or **Justification** (⑤), which uses oracle information. Different queries yield only slight differences in performance and none of them is statistically significant, even when ⑤ uses the human-written justification. We believe this is because we expand the retrieved text span by a context window (±150 words). As a result, this retrieval step does not need to be very precise to capture the relevant information.

Claim-focused Summarization We compare **zero-shot** (Ⓑ) and **few-shot** (⑥) prompts for generating the summary; **no summary** (⑦) directly feeds the text spans from second-stage retrieval to the veracity classifier. System ⑦ shows the worst performance across all metrics, suggesting that summarization matters. This may result from two primary factors: (1) The document length exceeds the context window capacity of DeBERTa, causing crucial information to be truncated. (2) our veracity classifier cannot easily discern the most relevant information given a large amount of context. Differences in the prompt (Ⓑ and ⑥) do not impact veracity classification results much but have differences under human inspection, which we discuss in the next section.

[9]Throughout our study, we use paired bootstrap tests for statistical significance between the results.

	Evidence Generation			Performance			
	FSR	SSR	Summary	Acc	Soft Acc	Macro-F1	MAE
		Claim only		25.5^{+}	68.0^{+}	27.5^{+}	1.12^{+}
		Claim + Justification		57.5^{+}	93.0^{+}	57.8^{+}	0.50^{+}
			Our Default System				
Ⓑ	subQs	subQs	`zero-shot-003`	33.0	74.5	34.5	0.99
			Ablation on first-stage retrieval				
①	Claim			24.5^{+}	71.5	18.0^{+}	1.15^{+}
②	Gold subQs			27.5	72.0	28.1^{+}	1.05^{+}
			Ablation on second-stage retrieval				
③		Claim		31.5	75.0	35.6	0.97
④		Gold subQs		31.5	73.0	35.4	1.03
⑤		Justification		33.0	71.5	37.2	1.01
			Ablation on summarization				
⑥			`few-shot-003`	35.0	76.5	36.2	0.94
⑦			no summary (raw doc)	29.0	66.0^{+}	26.3^{+}	1.18^{+}

Table 3: End-to-end fact-checking performance on the test set of CLAIMDECOMP. We ablate various stages of the model (FSR: first-stage retrieval; SSR: second-stage retrieval). Red indicates using oracle information. "+" denotes the result changes are statistically significant ($p < 0.05$) with respect to our default system.

5.3 Stability of First-stage Retrieval

As commercial search engines evolve over time, we conduct experiments to explore the reproducibility of our first-stage retrieval step. We use the default system setting in Table 3 and conducted three rounds of retrieval at $T = 0$, $T = 1$ week, and $T = 2$ months. We evaluate the Jaccard similarity of the sets of URLs retrieved from our queries to understand how much changes in the Bing API and the broader web change our results. We also evaluate the veracity of our system. Note that this Jaccard similarity is between the members of the URL sets (i.e., the URLs themselves), not capturing any lexical or domain similarity of the URLs.

Results are shown in Table 4. A noticeable trend is a decline in the Jaccard score between varying retrieval rounds over time. However, this decrease does not significantly impact the models' efficacy in the veracity assessment.

We caution that as the time gap increases, the set of documents retrieved from the Bing Search API could become considerably different, posing a challenge to consistently benchmark retrieval performance using commercial search engines. Therefore, we advocate for future research to focus on developing a comprehensive yet challenging document set that could be publicly released as a benchmark to spur research.

	Overlap	Acc	Soft-Acc	Ma-F1	MAE
Ours	-	33.0	74.5	34.5	0.99
1 week	0.48	33.5	74.0	36.8	0.98
2 months	0.30	29.5	73.5	32.3	1.03

Table 4: Model performance with respect to different rounds of retrieval at intervals of one week and two months. The overlap between "Ours" and subsequent document sets, measured with Jaccard score, decreases as the time gap increases. However, none of the changes in our downstream metrics is statistically significant.

6 Human Evaluation of Summaries

Summarizing documents from web search with large language models improves the performance of our fact-checking pipeline. However, these models can generate untruthful content (Bommasani et al., 2021; Chowdhery et al., 2022; Ouyang et al., 2022). Furthermore, as pointed out by Lim (2018), the accuracy of veracity classification alone does not entirely reflect the system's overall effectiveness, as certain labels such as "false" and "barely-true" may be ambiguous. We believe the true measure of our system's utility lies in the full package of summarized evidence it returns rather than just the accuracy of the veracity label. Therefore, we carry out two human studies, on comprehensiveness and faithfulness, to better understand intermediate outputs of the system.

Setting We randomly pick 50 claims which contain 200 document-summary pairs from the devel-

Summ-type	F	Minor	Major	NF	Avg score
zero-shot-001	65.8%	9.2%	20.0%	5.0%	3.45
zero-shot-003	66.0%	18.0%	16.0%	0.0%	3.50
few-shot-003	82.5%	6.5%	8.5%	2.5%	3.69

Table 5: Faithfulness Human Evaluation ($N = 200$). "F" denotes that the summary is factual and "NF" denotes that the summary is completely wrong. Few-shot prompting helps the model make fewer factual errors.

opment set of CLAIMDECOMP and run two human evaluation studies on this set. For each task, we recruited annotators from Amazon Mechanical Turk with a qualification test. In total, we recruited 17 worker for the faithfulness study and 15 workers for the comprehensiveness study. The details about crowdsourcing can be found in Appendix C.

Comparison Systems We compare the summaries generated from two prompts, `zero-shot-003` and `few-shot-003`, on GPT-3.5 (davinci-003). For the faithfulness study, we also compare the summaries generated through with zero-shot prompt on an earlier GPT model (davinci-001) (`zero-shot-001`) to see how the faithfulness varies for different models.

6.1 Faithfulness Evaluation

Goal We assess the frequency and degree to which the language model generates untruthful content during query-focused summarization. For each document and summary pair, annotators choose one of four labels below (see appendix C.1 for examples):

- **Faithful:** the summary accurately represents the meaning and details of the original document.
- **Minor Factual Error:** some details are not aligned with the original document, but the overall message remains intact.
- **Major Factual Error:** there are factual errors that result in the summary misrepresenting the original document.
- **Completely Wrong:** the language model hallucinates content that completely alters the meaning of the original document.

In addition to selecting a label, we ask annotators to provide a natural language justification for their choices. The annotations agree with a Fleiss Kappa score of 0.30. While this number is somewhat low, when we evaluated their justifications and we find many of the disagreements are because of subjectivity on the extent of factual error. We compute a consensus annotation via majority vote. We assign numerical scores to each label, where "Faithful", "Minor", "Major", and "Completely Wrong" correspond to 4, 3, 2, and 1 respectively and report average values. If all annotators disagree, we compute the average score and return the label that is nearest to the average score as a consensus.

Results The results are shown in Table 5. We see that **few-shot prompting substantially decreases the chance of hallucinations in the summaries**. When combining "Factual" and "Minor", we see 89% of the summaries are good enough to be used as evidence for the classifier. Additionally, by checking the unfaithful summaries, we find that they do not consist of useful hallucination like making a veracity judgment based on the parametric knowledge. Comparing the performance of `zero-shot-001` and `zero-shot-003`, we find that the weaker model makes more major factual errors. Together, they indicate that with stronger models and better prompts, we may expect these summarization models to improve further.

6.2 Comprehensiveness Evaluation

Goal We aim to measure the extent to which the claim-focused summaries are able to address the claim. This is subjective and difficult task to evaluate. Here, we leverage the human-annotated yes/no subquestions from CLAIMDECOMP as a proxy for evaluating the comprehensiveness of our summaries: if provided summary can help humans to answer more of these yes/no questions, we deem the summary to be more comprehensive.

In this task, annotators are given a summary / subquestion pair and label subquestion as "answerable", "partially answerable",[10] or "unanswerable", and additionally provide yes/no answer if the question is labeled as "answerable". Annotators were also asked to provide natural language justification for their answers. We collect this annotation on 161 questions associated with 50 claims. The annotations agree with a Fleiss Kappa score of 0.32.

Results The results are presented in Table 6. We see that zero-shot summaries yield more answerable questions than few-shot summaries. However, faithfulness evaluation hints that this is caused by hallucinations in zero-shot summaries; the system

[10] Sometimes the questions cannot be directly answered but can be inferred from the content of the summaries, or the summary at least contains relevant information. In such cases, we ask annotators to choose "partially answerable".

Summ-type	Ans	Partially Ans	UnAns
`zero-shot-003`	47.8%	22.4%	29.8%
`few-shot-003`	42.9%	21.1%	36.0%

Table 6: Human evaluation results on 161 subquestions from the same 50 claims we picked for the human study on faithfulness. "Ans", "Partially Ans", and "UnAns" denote the number of questions that are answerable, partially answerable, and unanswerable.

	Faithful	Minor	Unfaithful	Total
Ans	4	2	0	6
Partially Ans	6	1	1	8
Partially UnAns	13	5	11	30
UnAns	5	1	0	6
Total	28	10	12	50

Table 7: **Claim-level** statistics of `few-shot-003` taking both faithfulness and comprehensiveness into consideration. "Unfaithful" label aggregates "Major Error" and "Completely Wrong" labels. The claim-level labels are derived from the sub-parts as defined in section 6.3.

imputes information that seems to help, but which is not supported by the document.

Nevertheless, the few-shot summaries allow us to partially address over 60% of the **gold annotated** subquestions derived from the PolitiFact justification. We find this result encouraging: even though the system does not have access to these (often subtle) factors, it can retrieve information to enable a human annotator to make a judgment about them.

6.3 Combined Evaluation

While in previous section we evaluated faithfulness and comprehensiveness separately, here we conduct a **claim-level** evaluation: how many claims can be comprehensively addressed with a set of faithful summaries? We label a claim as **answerable** if all of its subquestions are answerable. If all subquestions are unanswerable the claim is **unanswerable**. Otherwise, we label claim as **partially unanswerable**. For claim-level faithfulness, we apply the same principles: a claim is faithful is all summaries are faithful, otherwise it is either unfaithful or contains minor factual errors. Table 7 shows the results by combining the two factors. We see that addressing every aspect of complex claims is still challenging: 36 out of 50 claims contain at least one unanswerable question. For claims that can be fully addressed (all questions are either answerable or partially answerable), only 1 out of 14 contains a major factual error in the summary.

7 Related Work

Retrieval augmented models Prior work has shown that a variety of NLP tasks could benefit from incorporating a retrieval component. Such tasks mainly include question answering (Chen et al., 2017; Kwiatkowski et al., 2019; Karpukhin et al., 2020; Khattab et al., 2021; Nakano et al., 2021), text generation (Lewis et al., 2020; Shi et al., 2023; Ram et al., 2023), language modeling (Guu et al., 2020; Khandelwal et al., 2020; Zhong et al., 2022), and dialog (Moghe et al., 2018; Fan et al., 2021; Thoppilan et al., 2022).

Most of these work assume having access to a fixed corpus, however, for the task of real-world fact-checking, no such corpus exists. In this work, we follow WebGPT (Nakano et al., 2021) and use Bing Search API to retrieve evidence from the wild web. Recent LLM agents such as Bing Chat and Google Bard follow this paradigm, so we believe these directions will be relevant for future work.

Question decomposition has been shown to be effective in evidence retrieval and question understanding for complex question answering (Talmor and Berant, 2018; Min et al., 2019; Qi et al., 2019; Perez et al., 2020; Wolfson et al., 2020; Geva et al., 2021). Question generation has also been shown to play a useful role in retrieval pipelines in open-domain QA (Sachan et al., 2022). In more recent research, it was demonstrated by Chen et al. (2022a) that such decompositions can also aid in retrieving evidence to assess complex claims and make veracity judgment. This observation is consistent with concurrent studies on fact-checking text generation outputs (Gao et al., 2022; Chen et al., 2022b; Liu et al., 2022) and Wikipedia (Kamoi et al., 2023).

8 Conclusion

We introduce a pipeline for realistic, automated fact-checking of complex political claims by retrieving raw evidence from web documents, improving final fact checking accuracy by integrating retrieved evidence. Our pipeline show promising results on the CLAIMDECOMP dataset. Yet, web search often cannot surface all the pieces of information necessary to verify a given claim. This work emphasizes the challenges of evidence retrieval in real-world scenarios and underscores the need for a human-in-the-loop fact-checking system.

Limitations and Future Directions

Performance is bottlenecked by the first-stage retrieval. The results in the last section show that 36.0% of questions are unanswerable using our most faithful claim-focused summaries. By investigating the unanswerable cases, we see that the following cases lead to retrieval failure: (1) no relevant information is available on the web except the fact-checking websites. These claims can be onerous to check, such as requiring talking to or emailing specific people to check facts. Those cases are beyond the scope of this work and we think a system doing triage for the claims, would be promising for future work. (2) No relevant subquestions are generated or the subquestions are not well decontextualized (Choi et al., 2021). In such cases, a stronger question generation model or decontextualization model can help further.

The need of human-in-the-loop fact-checking. To address the failures in the first-stage retrieval and the potential errors in the summarization stage, we envision a human-in-the-loop fact-checking system. This system begins with the automated pipeline presented in this paper, which provides fact-checkers with summarized documents and judgments. If the fact-checkers deem these documents unsatisfactory, the system reveals the subquestions used for evidence retrieval, allowing fact-checkers to rerun the search. The system then retrieves additional documents and generates updated summaries. This iterative process continues until the fact-checkers are satisfied with the retrieved evidence. Moreover, the system could further learn from the fact-check feedback to improve itself: for example, the system could learn what questions are important to retrieve good evidence and what questions are not according to the fact-checker. In general, we believe such systems will be necessary, but developing them is outside of the scope of this work.

Scope of facts checked. Our work only addresses English-language political claims. Misinformation in other languages is a crucial problem that we believe future work should address. Moreover, even within English, there is a strong need for fact-checking systems that can address other kinds of claims that have a different distribution; for example, claims from social media, which are often embedded in images or memes. Nevertheless, we believe the decomposition and retrieval approach here can play a role in such systems as well.

Acknowledgments

This work was partially supported by NSF CAREER Award IIS-2145280, by Good Systems,[11] a UT Austin Grand Challenge to develop responsible AI technologies, and by grants from Salesforce Inc. and Open Philanthropy. We thank the UT Austin NLP community for feedback on the earlier drafts of the paper.

References

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.

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.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Erik Brand, Kevin Roitero, Michael Soprano, Afshin Rahimi, and Gianluca Demartini. 2022. A neural model to jointly predict and explain truthfulness of statements. *ACM Journal of Data and Information Quality*, 15(1):1–19.

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

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.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022a. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. 2022b. PropSegmEnt: A Large-Scale Corpus for Proposition-Level Segmentation and Entailment Recognition. *arXiv eprint arxiv:2212.10750*.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with KNN-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

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*. ACL.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, N. Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2022. Rarr: Researching and revising what language models say, using language models.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *arXiv eprint arxiv:2209.12356*.

Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-Checking. Technical report, Reuters Institute, University of Oxford.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. *arXiv eprint arxiv:2303.01432*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language

models. In *International Conference on Learning Representations*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics*, 9:929–944.

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

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chloe Lim. 2018. Checking how fact-checkers check. *Research & Politics*, 5(3):2053168018786848.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2022. Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. *arXiv eprint arxiv:2212.07981*.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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)*.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, Online. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. 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 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

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.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 36.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. *arXiv preprint arXiv:2301.12652*.

Prakhar Singh, Anubrata Das, Junyi Jessy Li, and Matthew Lease. 2021. The case for claim difficulty assessment in automatic fact checking. *arXiv preprint arXiv:2109.09689*.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex 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 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language Models for Dialog Applications. *arXiv preprint arXiv:2201.08239*.

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.

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.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *arXiv eprint arxiv:2301.13848*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5657–5673, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Appendix

A Experimental Details

A.1 List of Websites being Filtered

- `www.politifact.com`
- `www.snopes.com`
- `www.factcheck.org`
- `www.washingtonpost.com/news/fact-checker/`
- `www.apnews.com/hub/ap-fact-check`
- `www.fullfact.org`
- `www.reuters.com/fact-check`

We also filter the URLs that contain "fact-check" or "factcheck"; we also filter any PDF files and videos.

Claim: Viral image stated on June 8, 2020 in post on Facebook: Cops in Norway: require 3 years of training, 4 people killed since 2002. Cops in Finland: require 2 years of training, 7 people killed since 2000. Cops in Iceland: require 2 years of training, 1 person killed since ever. Cops in the U.S.: require 21 weeks of training, 8,000+ people killed since 2001.

Suppose you are a fact-checker, generate several yes or no questions to help me answer if this claim is true or false.

Questions:
Does Norway require 3 years of training for cops?
Have Norwegian cops killed 4 people since the early 2000's?
Does Finland require only 2 years of training for police?
Have Finnish police killed 7 people since 2000?
Does Iceland only require 2 years of training for cops?
Have Iceland cops only killed 1 person ever?
Does the U.S. require only 21 weeks of training for cops?
Have U.S. cops killed more than 8,000 people since 2001?
Do experts associate only training time with police-related shooting fatalities?

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.

Suppose you are a fact-checker, generate several yes or no questions to help me answer if this claim is true or false.

Questions:
Is the U.S. the only major oil-producing nation to ban exports of crude oil?
Is the self-imposed ban on crude oil export of U.S a complete ban?

Claim: William Barr stated on September 2, 2020 in a CNN interview: We indicted someone in Texas, 1,700 ballots collected from people who could vote, he made them out and voted for the person he wanted to.

Suppose you are a fact-checker, generate several yes or no questions to help me answer if this claim is true or false.

Questions:
Were 1700 mail-in ballots investigated for fraud in Texas during the 2020 election?
Did the Justice Department indict someone in Texas for voter fraud?
Did widespread mail-in order fraud happen in Texas during the 2020 election?
Did voter disenfranchisement happen in Texas during the 2020 election?

Claim: INPUT CLAIM
Suppose you are a fact-checker, generate several yes or no questions to help me answer if this claim is true or false.

Questions:
MODEL OUTPUT

Figure 6: Few-shot prompt we used to generate sub-questions in this paper.

A.2 Question Generation Prompt and Deduplication

The prompt we used to generate the questions is shown in Figure 6. Since the generated question set sometimes contains duplicates, we delete the duplicated questions according to the exact string match.

A.3 Question-focused Summarization Prompt

The zero-shot and few-shot prompts we used to generate the claim-focused summaries are shown in Figure 7 and Figure 8 respectively.

A.4 Hyperparameters of Veracity Classifier

- Model: DeBERTa-large
- Batch size: 32

Suppose you are assisting a fact-checker to fact-check the claim: INPUT CLAIM

Summarize the relevant information from the document in 1-2 sentences. Your response should provide a clear and concise summary of the relevant information contained in the document. Do not include a judgment about the claim and do not repeat any information from the claim that is not supported by the document.

Summarization:
MODEL OUTPUT

Figure 7: Zero-shot prompt we used to generate the claim-focused summaries in this paper.

	First-stage	Second-stage	Summ
# documents	45.0	7.7	4.0
# words	70,245	2,710	251

Table 8: Average number of unique documents and average number of words in total from those documents after each stage of our pipeline.

- Max sequence length: 512
- Epochs: 25
- Initial learning rate: 3e-5
- Optimizer: Adam with linear decay
- Metric for selecting best dev model: MAE
- Random seed of 5 runs: 290032, 33432, 7876, 366, 77
- Training device: NVIDIA-A6000

B Information Compression through the Pipeline

Our pipeline progressively refines the crucial data needed to validate a claim. Table 8 demonstrates the average count of unique documents and the total word count in these documents after each phase of our pipeline under both temporal and site constraints.

C Human Study

C.1 Examples of Unfaithful Summaries

Figure 12 shows three examples containing unfaithful content. We see that the "Minor" error does not affect the interpretation of the original document while "Major" and "Completely Wrong" errors alter the view.

Document: Vote by mail: Which states allow absentee voting - Washington Post
Content: excuse to vote absentee and states that will allow fear of the coronavirus as an excuse . In response to the coronavirus , nearly half of all states expanded access to mail ballots for their primaries , either by allowing fear of the coronavirus as a reason or proactively sending an application or ballot to every registered voter . Fewer have taken action for the general election , as the move has become increasingly partisan and subject to litigation . President Trump has made numerous unfounded claims that mail-in voting will create widespread abuse and fraud . His suspicions are out of step with the views of election experts and many within his own party , who are building large-scale vote-by-mail programs . A recent analysis by The Washington Post found only 372 cases of potential fraud out of roughly 14.6 million ballots cast by mail in 2016 and 2018 . [Examining the arguments against voting by mail : Does it really lead to fraud or benefit only Democrats ?] Only a quarter of voters used mailed ballots in 2018 , and they mostly resided in a handful of states . Nearly everyone who voted in Oregon , the first state to issue all ballots by mail in 2000 , did so by mail . But in most states , fewer than 10 percent of voters did . # # # # Most places expanding vote-by-mail in November had limited mail-in voting in 2018 Percentage of votes cast by mail in 2018 midterm elections Even in states that havenât made absentee voting easier , the number of ballot requests is still expected to spike . To meet this challenge , local election officials will have to overcome numerous hurdles with little time and money to spare . They must acquire large volumes of specialized envelopes and paper . Additional staff , and in some cases machines , are necessary to open , sort and tabulate postal ballots and verify

Suppose you are assisting a fact-checker to fact-check the claim:
"Donald Trump stated on April 7, 2020 in a press briefing: With voting by mail, “you get thousands and thousands of people sitting in somebody's living room, signing ballots all over the place.”"
Summarize the relevant information from the document in 1-2 sentences. Your response should provide a clear and concise summary of the relevant information contained in the document. Do not include a judgment about the claim and do not repeat any information from the claim that is not supported by the document:

Trump’s suspicions are out of step with the views of election experts and many within his own party, who are building large-scale vote-by-mail programs. The Washington Post conducted an analysis that found only 372 cases of potential fraud out of approximately 14.6 million ballots cast by mail in the 2016 and 2018 elections.

Document: Governor Ron DeSantis' State of the State Address - CBS Miami
Content: hard and one thing is clear : we have a great First Lady . I 'm convinced that this whole Governor thing is just a way for the people of Florida to have Casey as their First Lady . Our kids Madison and Mason have taken over the Governor 's mansion but the baby-proofing has been successful – I can report that no artifacts of Florida history have yet to be destroyed . But stay tuned – Mason will be walking very soon . Just last week , Casey traveled to NW Florida to survey the Hurricane Michael recovery efforts . We have both been to the region several times over the past few months . To the people of NW Florida : we pledged to stand with you as you work to rebuild your communities stronger than before – and we have followed through on that pledge . I 've already traveled to Washington , DC and secured a historic commitment from the Trump administration to provide assistance to the communities that Michael battered . My administration has worked to cut through red tape to expedite relief efforts and , thanks to the leadership of your former colleague Jared Moskowitz , we are making great progress . Here with us today is Mayor Al Cathey and Administrator Tanya Castro from Mexico Beach , which was ground zero for Hurricane Michael 's wrath . They and many others affected by the storm have displayed remarkable resilience in the face of huge obstacles . They deserve our admiration – and our support . These communities will not be rebuilt in days , weeks or months , but they will be rebuilt . They will be rebuilt because we as Floridians will answer the call as we always do . God bless the good people of NW Florida . Executive energy and leadership are necessary to meet fully the challenges that are before us – but they are not sufficient . In a

Suppose you are assisting a fact-checker to fact-check the claim:
“Ron DeSantis stated on March 5, 2019 in his State of the State address: "I’ve already traveled to Washington, D.C., and secured a historic commitment from the Trump administration to provide assistance to the communities that Michael battered."
Summarize the relevant information from the document in 1-2 sentences. Your response should provide a clear and concise summary of the relevant information contained in the document. Do not include a judgment about the claim and do not repeat any information from the claim that is not supported by the document:
Ron DeSantis did state in his address that he traveled to Washington, D.C. and secured a commitment from the Trump administration. However, no additional evidence is provided about this claim.

Document: Cuban government unveils controversial cybersecurity regulation
Content: others . Activists , opponents , unofficial organizations and Internet users fear that the new regulations will be applied at the discretion of the government and will serve to silence dissident voices on the networks and prosecute those who criticize the Cuban system and its leaders . In this sense , complaints and criticisms of the island ’ s authorities due to the new regulation have already started being published on the social networks . In addition , in this Tuesday ’ s Gaceta Oficial other regulations were published that , according to the Ministry of Communications , “ update ” and “ strengthen ” Cuba ’ s legal body on these issues . Among these is Decree-Law No . 35 on “ Telecommunications , Information and Communication Technologies and the Use of the Radioelectric Spectrum , ” which establishes that “ the Cuban state is the owner of public telecommunications services and has the power to regulate , order , control and supervise the different telecommunications/ICT services and networks ; as well as granting the right to its exploitation , ” Cubadebate pointed out . Finally , the package includes a group of regulations on Information and Communication Technologies , the Radioelectric Spectrum , the use of Radiocommunication Services by Satellites and Interconnection , Access and Essential Installations of Telecommunication Networks . OnCuba Staff

Suppose you are assisting a fact-checker to fact-check the claim:
“Marco Rubio stated on December 17, 2014: "The reason why Cubans don't have access to 21st century telecommunications — like smart phones, like access to the Internet — is because it is illegal in Cuba."”
Summarize the relevant information from the document in 1-2 sentences. Your response should provide a clear and concise summary of the relevant information contained in the document. Do not include a judgment about the claim and do not repeat any information from the claim that is not supported by the document:
Cuba is passing new regulations on telecommunication, establishing that the Cuban state can regulate and supervise different telecommunication services and networks. This includes Decree-Law No . 35 on “ Telecommunications , Information and Communication Technologies and the Use of the Radioelectric Spectrum.”

Document: FBI Chief Claims 'Gaps' Remain In Syrian Refugee Screening Process
Content: # # # # Published : Fri , Oct 9th 2015 @ 1:13 pm EDT FBI Director James Comey voiced his concern during a Senate Homeland Security Committee hearing about the lack of information available to screen Syrian refugees coming into the United States . “ My concern there is there are certain gaps ... in the data available to us , ” Comey said . “ There is risk associated of bringing anybody in from the outside , but specifically from a conflict zone like that , ” he added . This hearing comes after the Obama administration ’ s announcement to allow 85,000 refugees into the U.S. next year and 100,000 in 2017 . Those are huge increases from the current level of 70,000 . The Obama administration has said that at least 10,000 of the refugees next year will come from Syria . Many lawmakers are concerned that ISIS extremists or other terrorists could make it into the U.S. as a refugee due to the lack of intelligence and information available about the area . “ The intelligence that we have of this particular conflict zone is not as rich as we would like it to be , ” admitted Nicholas Rasmussen , the head of the National Counterterrorism Center during the hearing . Even Homeland Security Secretary Jeh Johnson expressed concern saying , “ We should do the right thing by accepting more , but we should be careful in doing it. ” Currently the U.S. has taken in around 2,000 Syrian refugees and has contributed about $ 4 billion in foreign aid to the Syrian refugee crisis , more than any other country . Read more at The Hill . refugees

Suppose you are assisting a fact-checker to fact-check the claim:
“Rob Portman stated on October 20, 2016 in an Ohio Senate debate in Cleveland: "The director of the FBI said that we cannot figure out who (Syrian refugees) are, what their intentions are, because we have no information on them, because we don’t have any contact with the Syrian government or any people on the ground to be able to determine that.""
Summarize the relevant information from the document in 1-2 sentences. Your response should provide a clear and concise summary of the relevant information contained in the document. Do not include a judgment about the claim and do not repeat any information from the claim that is not supported by the document:
FBI Director James Comey voiced concern during a Senate Homeland Security Committee hearing about the lack of information available to screen Syrian refugees coming into the United States. Nicholas Rasmussen says we do not have as much information about Syria as we would like and Homeland Security Secretary Jeh Johnson says we need to be careful.

Figure 8: Few-shot prompt we used to generate the claim-focused summaries in this paper.

C.2 Recruiting Process

Faithfulness study We set up a qualification test that consists of 5 examples. We selected workers from MTurk if they get more than 3/5 examples correct according to our curated labels and if they write reasonable rationales. In total, there are 31 workers who took the qualification test and we selected 15 of them for the task. We pay $3 for the qualification test and $2 dollars for one HIT that contains 4 document-summary pairs in the actual task. The detailed instructions and the annotation interface is shown in Figure 10.

Comprehensiveness study We set up a qualification test that consists of 10 examples. We selected workers from MTurk if they got more than 7/10 questions right according to our curated labels and if they write reasonable rationales. In total, there are 28 workers who took the qualification test and we selected 17 of them for the task. We pay $3 for the qualification test and $0.3 dollars for one

```
Use summary of documents to score how likely this claim is
true at the scale of 0 to 100, 0 being pants-on-fire
and 100 being true.

Explain your reasoning first and output your predicted score.

Claim to fact-check: INPUT CLAIM

Summary to use:
Document 0: INPUT SUMMARY 0
Document 1: INPUT SUMMARY 1
Document 2: INPUT SUMMARY 2
Document 3: INPUT SUMMARY 3

Format your output like this
Explanation: Your explanation
Score: Your prediction
```

Figure 9: Zero-shot prompt for Claim + summary

question in the actual task.

The detailed instructions and the annotation interface is shown in Figure 11.

D Using LLMs as a Veracity Classifier

We experiment with using ChatGPT (`gpt-3.5-turbo`) as the classifier in the final stage. Since ChatGPT is not trained on our training set, it does not have access to the label distribution of the dataset. To make a fair comparison with the DeBERTa model, instead of directly predicting a discrete label (one out of the six labels), we prompt the model to explain its reasoning process and predict a truthfulness score on a scale of 0 to 100, 0 for the claim being false and 100 for true. We then rank the examples according to the predicted scores and map the scores to discrete labels to the label distribution of the training set. To be specific, we rank the examples in the training set by their labels, assigning the lowest rank to *pants-on-fire* and the highest to *true*. Each label, denoted as l_i, corresponds to a percentile p_i. We then map the predicted score falling between p_i and p_{i+1} to the label l_i. We use a zero-shot prompt[12] to produce the score and the prompt is shown in Figure 9.

The results are shown in Table 9. Comparing the **claim-only** results from the two models, we see that ChatGPT achieves slightly better performance than DeBERTa. However, unlike the DeBERTa model, when adding the summary, we see a notable performance drop for ChatGPT. We argue that this might be because ChatGPT relies heavily on prior knowledge and it is not able to use the provided summary effectively. We believe improving this is a promising direction for future work.

[12]We also experimented with few-shot prompts. However, these did not yield better performance than the zero-shot prompt.

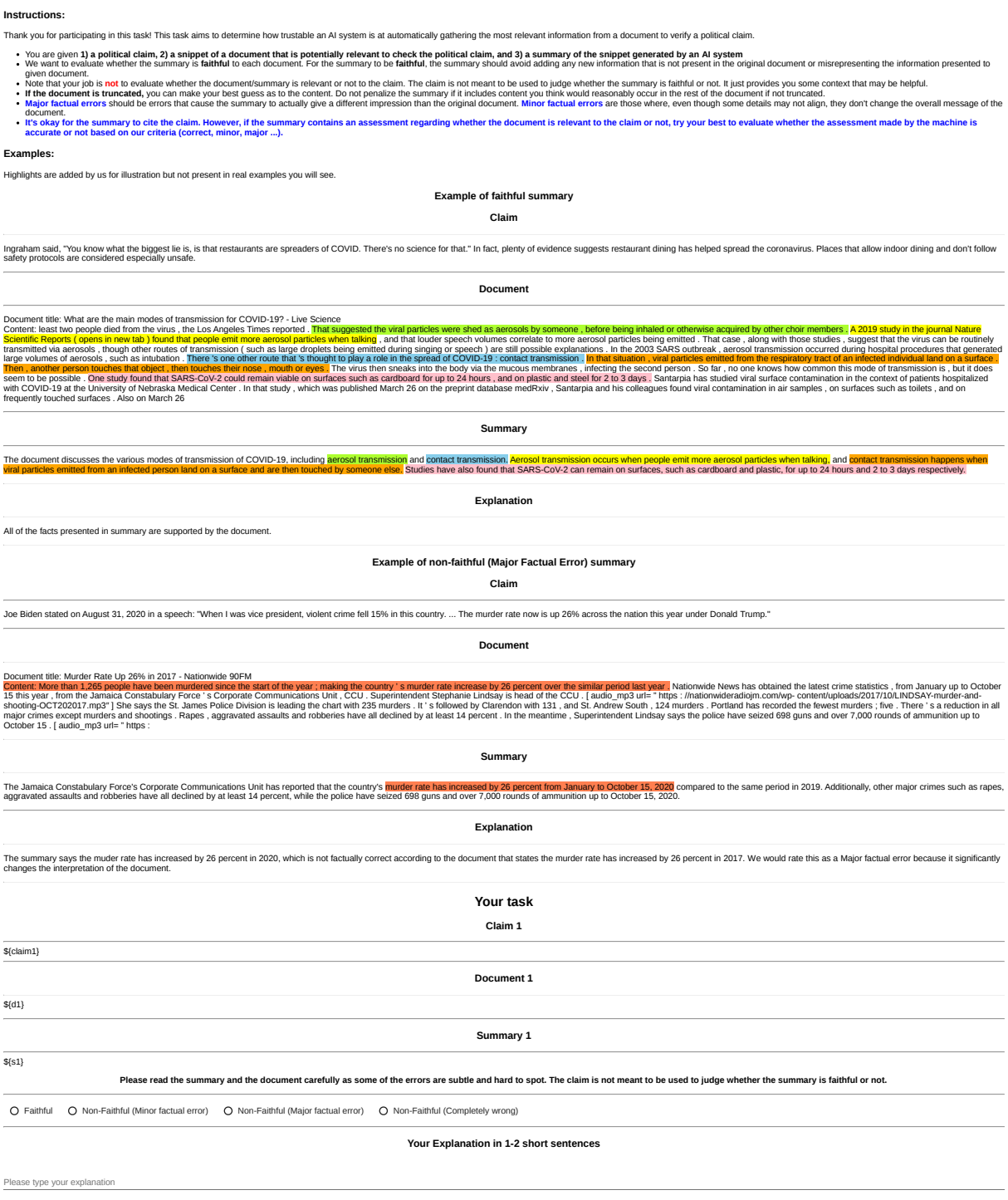
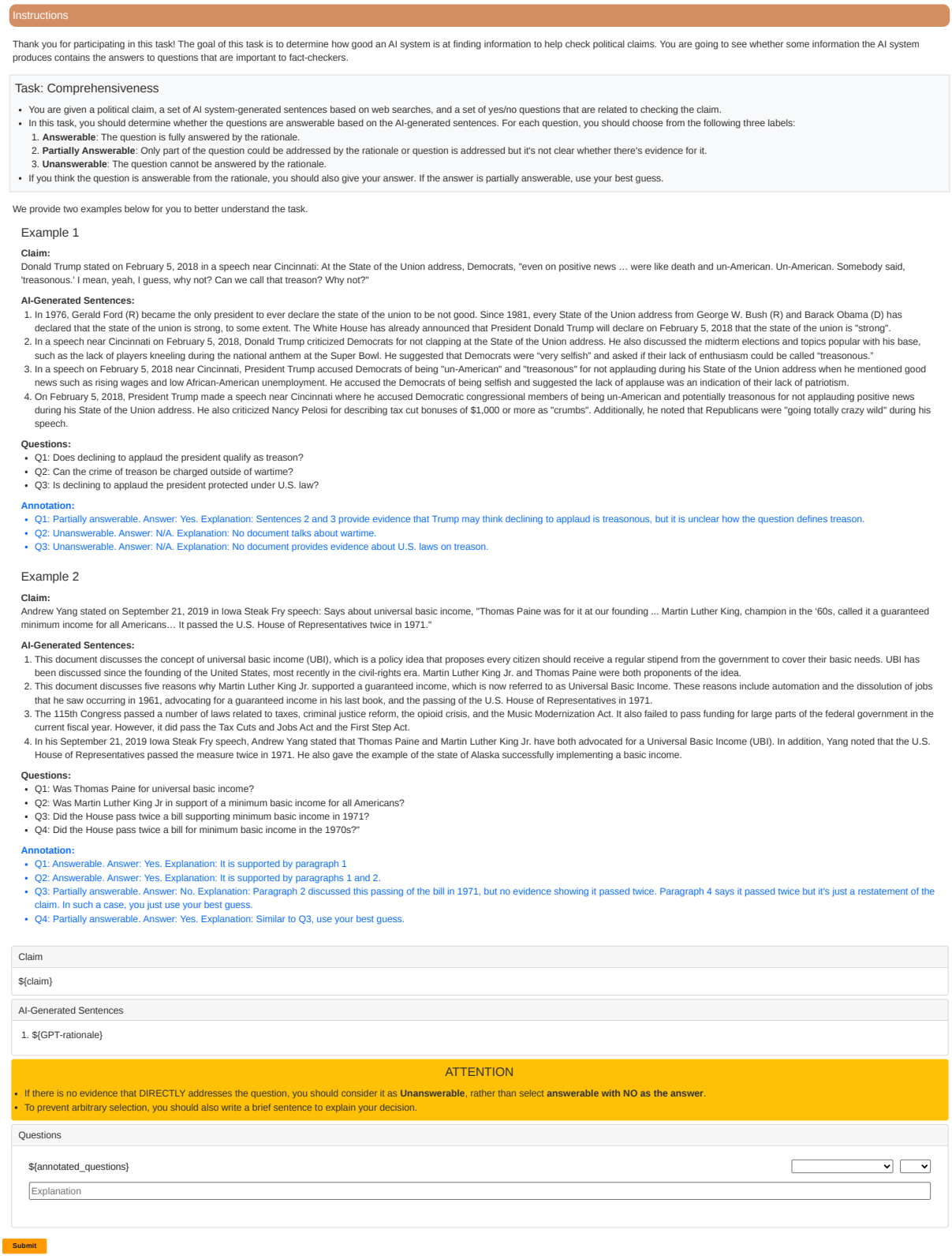


Figure 10: Interface of the faithfulness study we conducted in Section 6.1.

Model	Evidence	Acc	Soft-Acc	Macro-F1	MAE
ChatGPT	Claim only	32.0	66.0	31.0	1.16
	Claim + summary	24.5	67.5	25.7	1.25
DeBERTa-large	Claim only	25.5	68.0	27.5	1.12
	Claim + summary	33.0	74.5	34.5	0.99

Table 9: Veracity classification performance on the test set of CLAIMDECOMP with different prompts using ChatGPT.



Instructions

Thank you for participating in this task! The goal of this task is to determine how good an AI system is at finding information to help check political claims. You are going to see whether some information the AI system produces contains the answers to questions that are important to fact-checkers.

Task: Comprehensiveness

- You are given a political claim, a set of AI system-generated sentences based on web searches, and a set of yes/no questions that are related to checking the claim.
- In this task, you should determine whether the questions are answerable based on the AI-generated sentences. For each question, you should choose from the following three labels:
  1. **Answerable**: The question is fully answered by the rationale.
  2. **Partially Answerable**: Only part of the question could be addressed by the rationale or question is addressed but it's not clear whether there's evidence for it.
  3. **Unanswerable**: The question cannot be answered by the rationale.
- If you think the question is answerable from the rationale, you should also give your answer. If the answer is partially answerable, use your best guess.

We provide two examples below for you to better understand the task.

Example 1

Claim:
Donald Trump stated on February 5, 2018 in a speech near Cincinnati: At the State of the Union address, Democrats, "even on positive news … were like death and un-American. Un-American. Somebody said, 'treasonous.' I mean, yeah, I guess, why not? Can we call that treason? Why not?"

AI-Generated Sentences:

1. In 1976, Gerald Ford (R) became the only president to ever declare the state of the union to be not good. Since 1981, every State of the Union address from George W. Bush (R) and Barack Obama (D) has declared that the state of the union is strong, to some extent. The White House has already announced that President Donald Trump will declare on February 5, 2018 that the state of the union is "strong".
2. In a speech near Cincinnati on February 5, 2018, Donald Trump criticized Democrats for not clapping at the State of the Union address. He also discussed the midterm elections and topics popular with his base, such as the lack of players kneeling during the national anthem at the Super Bowl. He suggested that Democrats were "very selfish" and asked if their lack of enthusiasm could be called "treasonous."
3. In a speech on February 5, 2018 near Cincinnati, President Trump accused Democrats of being "un-American" and "treasonous" for not applauding during his State of the Union address when he mentioned good news such as rising wages and low African-American unemployment. He accused the Democrats of being selfish and suggested the lack of applause was an indication of their lack of patriotism.
4. On February 5, 2018, President Trump made a speech near Cincinnati where he accused Democratic congressional members of being un-American and potentially treasonous for not applauding positive news during his State of the Union address. He also criticized Nancy Pelosi for describing tax cut bonuses of $1,000 or more as "crumbs". Additionally, he noted that Republicans were "going totally crazy wild" during his speech.

Questions:

- Q1: Does declining to applaud the president qualify as treason?
- Q2: Can the crime of treason be charged outside of wartime?
- Q3: Is declining to applaud the president protected under U.S. law?

Annotation:

- Q1: Partially answerable. Answer: Yes. Explanation: Sentences 2 and 3 provide evidence that Trump may think declining to applaud is treasonous, but it is unclear how the question defines treason.
- Q2: Unanswerable. Answer: N/A. Explanation: No document talks about wartime.
- Q3: Unanswerable. Answer: N/A. Explanation: No document provides evidence about U.S. laws on treason.

Example 2

Claim:
Andrew Yang stated on September 21, 2019 in Iowa Steak Fry speech: Says about universal basic income, "Thomas Paine was for it at our founding ... Martin Luther King, champion in the '60s, called it a guaranteed minimum income for all Americans… It passed the U.S. House of Representatives twice in 1971."

AI-Generated Sentences:

1. This document discusses the concept of universal basic income (UBI), which is a policy idea that proposes every citizen should receive a regular stipend from the government to cover their basic needs. UBI has been discussed since the founding of the United States, most recently in the civil-rights era. Martin Luther King Jr. and Thomas Paine were both proponents of the idea.
2. This document discusses five reasons why Martin Luther King Jr. supported a guaranteed income, which is now referred to as Universal Basic Income. These reasons include automation and the dissolution of jobs that he saw occurring in 1961, advocating for a guaranteed income in his last book, and the passing of the U.S. House of Representatives in 1971.
3. The 115th Congress passed a number of laws related to taxes, criminal justice reform, the opioid crisis, and the Music Modernization Act. It also failed to pass funding for large parts of the federal government in the current fiscal year. However, it did pass the Tax Cuts and Jobs Act and the First Step Act.
4. In his September 21, 2019 Iowa Steak Fry speech, Andrew Yang stated that Thomas Paine and Martin Luther King Jr. have both advocated for a Universal Basic Income (UBI). In addition, Yang noted that the U.S. House of Representatives passed the measure twice in 1971. He also gave the example of the state of Alaska successfully implementing a basic income.

Questions:

- Q1: Was Thomas Paine for universal basic income?
- Q2: Was Martin Luther King Jr in support of a minimum basic income for all Americans?
- Q3: Did the House pass twice a bill supporting minimum basic income in 1971?
- Q4: Did the House pass twice a bill for minimum basic income in the 1970s?"

Annotation:

- Q1: Answerable. Answer: Yes. Explanation: It is supported by paragraph 1
- Q2: Answerable. Answer: Yes. Explanation: It is supported by paragraphs 1 and 2.
- Q3: Partially answerable. Answer: No. Explanation: Paragraph 2 discussed this passing of the bill in 1971, but no evidence showing it passed twice. Paragraph 4 says it passed twice but it's just a restatement of the claim. In such a case, you just use your best guess.
- Q4: Partially answerable. Answer: Yes. Explanation: Similar to Q3, use your best guess.

Claim

${claim}

AI-Generated Sentences

1. ${GPT-rationale}

ATTENTION

- If there is no evidence that DIRECTLY addresses the question, you should consider it as **Unanswerable**, rather than select **answerable with NO as the answer**.
- To prevent arbitrary selection, you should also write a brief sentence to explain your decision.

Questions

${annotated_questions}

Explanation

Submit

Figure 11: Interface of the comprehensiveness study we conducted in Section 6.2.

Claim: Facebook posts stated on September 29, 2021: The White House “created a fake set for (President Joe) Biden to get his booster shot. The entire Biden presidency is one giant charade.”
Document Title: President Joe Biden Gets COVID Booster Vaccine Shot - PEOPLE.com
Content: President Joe Biden received his COVID-19 booster shot .. Biden received his first dose of Pfizer-BioNTech novel coronavirus vaccine in December at Delaware 's ChristianaCare Hospital. His second shot was administered there three weeks later. Both shots were televised …
Summary: On Monday, President Joe Biden received his COVID-19 booster shot on camera, administered by a nurse with the White House Medical Unit. Biden has previously received both of his initial doses of the Pfizer-BioNTech novel coronavirus vaccine in December at Delaware's ChristianaCare Hospital.
Minor Error: The summaries says both shots are received in December while the document says the first one was in December and the second one was three weeks later. The summary slightly changes the information in the original document.

Claim: Elizabeth Warren stated on November 20, 2019 in a debate: "Today in America — a new study came out — 20 years out, whites who borrowed money, 94% of them have paid off their student loan debt; 5% of African Americans have paid it off.”
Document Title: More Evidence Student Borrowers Prepay Their Loans - Forbes
Content: Most borrowers will still fully pay off their loans before they are required to do so … Borrowers pay off almost all loans (94 %) with a final payment larger than the scheduled monthly payments they are required to make …
Summary: A new report from the Consumer Financial Protection Bureau found that 94% of borrowers pay off their student loans before the scheduled payoff date, typically doing so in five years or less ...
Major Error: The document says borrowers pay off 94% of the loan not 94% of borrowers pay off their loans

Claim: Andrew Giuliani stated on May 18, 2021 in a news conference: "The one good thing about the antibodies if you’ve had it, is it actually is even better than the vaccine, and here's why. With the vaccine you can still transmit, with the antibodies you can't transmit.”
Document Title: COVID-19: Long-term effects - Mayo Clinic
Content: It involves extreme fatigue that worsens with physical or mental activity , but doesn't improve with rest … What should you do if you have post-COVID-19 syndrome symptoms ? If you 're having symptoms of post-COVID-19 syndrome , talk to your health care provider …
Summary: The Centers for Disease Control and Prevention states that there is no evidence to suggest that people who have recovered from COVID-19 and have antibodies are not able to transmit the virus.
Completely Wrong: The document is about the long-term effects of COVID-19. However, model is likely utilizing its parameterized knowledge and draws the conclusion directly.

Figure 12: Three examples from the faithfulness evaluation (Section 6.1), showing the cases of minor error, major error, and completely wrong, respectively. Red text denotes the mismatches between the summary and the document.