FAVIQ: FAct Verification from Information-seeking Questions

Jungsoo Park[1*] **Sewon Min**[2*] **Jaewoo Kang**[1]
Luke Zettlemoyer[2] **Hannaneh Hajishirzi**[2,3]
[1]Korea University [2]University of Washington [3]Allen Institute of AI
`{jungsoo_park,kangj}@korea.ac.kr`
`{sewon,lsz,hannaneh}@cs.washington.edu`

Abstract

Despite significant interest in developing general purpose fact checking models, it is challenging to construct a large-scale fact verification dataset with realistic real-world claims. Existing claims are either authored by crowdworkers, thereby introducing subtle biases that are difficult to control for, or manually verified by professional fact checkers, causing them to be expensive and limited in scale. In this paper, we construct a large-scale challenging fact verification dataset called FAVIQ, consisting of 188k claims derived from an existing corpus of ambiguous information-seeking questions. The ambiguities in the questions enable automatically constructing true and false claims that reflect user confusions (e.g., the year of the movie being filmed vs. being released). Claims in FAVIQ are verified to be natural, contain little lexical bias, and require a complete understanding of the evidence for verification. Our experiments show that the state-of-the-art models are far from solving our new task. Moreover, training on our data helps in professional fact-checking, outperforming models trained on the widely used dataset FEVER or in-domain data by up to 17% absolute. Altogether, our data will serve as a challenging benchmark for natural language understanding and support future progress in professional fact checking.[1]

1 Introduction

Fact verification, the task of verifying the factuality of the natural language claim, is an important NLP application (Cohen et al., 2011) and has also been used to evaluate the amount of external knowledge a model has learned (Petroni et al., 2021). However, it is challenging to construct fact verification data with claims that contain realistic and implicit misinformation. Crowdsourced claims from prior work such as FEVER (Thorne et al., 2018a) are written with minimal edits to reference sentences, leading to strong lexical biases such as the overuse of explicit negation and unrealistic misinformation that is less likely to occur in real life (Schuster et al., 2019). On the other hand, data constructed by professional fact-checkers are expensive and are typically small-scale (Hanselowski et al., 2019).

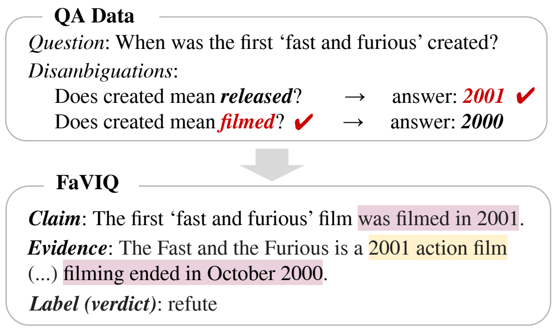

Figure 1: An example of a `refute` claim on FAVIQ, constructed using ambiguity in the information-seeking question, e.g., through a crossover of the year of the film *being released* and *being filmed*.

In this paper, we show it is possible to use information-seeking questions (Kwiatkowski et al., 2019) and their ambiguities (Min et al., 2020) to construct a large-scale, challenging, and realistic fact verification dataset. Information-seeking questions are inherently ambiguous because users do not know the answers to the questions they are posing. For example, in Figure 1, the question is ambiguous because the filming of the movie and the release of the movie can both be seen as the creation time.

We introduce a new dataset **FAVIQ—FAct Verification derived from Information-seeking Questions**, which uses such ambiguities to generate challenging fact verification problems. For instance, the claim in Figure 1 requires the model to identify that the movie released in 2001 is in fact filmed in 2000 and to return `refute`. Like this, claims generated through the crossover of the disambiguation of information-seeking questions

*Equal Contribution
[1]Data available at `https://faviq.github.io`.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5154 - 5166
May 22-27, 2022

are likely to contain misinformation that real users are easily confused with. We automatically generate such claims by composing valid and invalid question-answer pairs and transforming them into textual claims using a neural model. The data is further augmented by claims from regular question-answer annotations.

In total, FAVIQ consists of 188k claims. We manually verified a subset of claims to ensure that they are as natural as human-written claims. Our analysis shows that the claims have significantly lower lexical bias than existing crowdsourced claims; claims involve diverse types of distinct entities, events, or properties that are semantically close, being more realistic and harder to verify without a complete understanding of the evidence text.

Our experiments show that a model with no background knowledge performs only slightly better than random guessing, and the state-of-the-art model achieves an accuracy of 65%, leaving significant room for improvement. Furthermore, training on FAVIQ improves the accuracy of verification of claims written by professional fact checkers, outperforming models trained on the target data only or pretrained on FEVER by up to 17% absolute. Together, our experiments demonstrate that FAVIQ is a challenging benchmark as well as a useful resource for professional fact checking.

To summarize, our contributions are three-fold:

1. We introduce FAVIQ, a fact verification dataset consisting of 188k claims. By leveraging information-seeking questions and their natural ambiguities, our claims require the identification of entities, events, or properties that are semantically close but distinct, making the fact verification problem very challenging and realistic.
2. Our experiments show that the state-of-the-art fact verification models are far from solving FAVIQ, indicating significant room for improvement.
3. Training on FAVIQ significantly improves the verification of claims written by professional fact checkers, indicating that FAVIQ can support progress in professional fact checking.

2 Related Work

Fact verification Fact verification is crucial for real-world applications (Cohen et al., 2011) and as a benchmark to evaluate the knowledge in a model (Petroni et al., 2021).

One line of work has studied professional fact checking, dealing with claims collected by professional fact checkers in specific domains (Vlachos and Riedel, 2014; Ferreira and Vlachos, 2016; Augenstein et al., 2019; Hanselowski et al., 2019). While such data contains realistic claims that have occurred in the real world, it is expensive to construct as it requires labor from professional fact checkers. Moreover, it is less suitable as a benchmark due to lack of a standard evidence corpus such as Wikipedia[2] and ambiguities in labels.[3]

Other fact verification datasets are collected through crowdsourcing (e.g., FEVER (Thorne et al., 2018a) and its variants (Thorne et al., 2018b; Thorne and Vlachos, 2019)) by altering a word or negating the reference text to intentionally make true or false claims. This process leads to large-scale datasets but with strong artifacts and unrealistic claims (Schuster et al., 2019; Thorne and Vlachos, 2019; Eisenschlos et al., 2021). Consequently, a trivial claim-only baseline with no evidence achieves near 80% (Petroni et al. (2021), verified in Section 4.1). While more recent work proposes new crowdsourcing methods that alleviate artifacts (Schuster et al., 2021; Eisenschlos et al., 2021), their claims are still written given particular evidence text, being vulnerable to subtle lexical biases that can be hard to explicitly measure.

We construct a fact verification dataset from highly ambiguous information-seeking questions. Our claims have significantly less lexical bias than other crowdsourced ones (Figure 3), contain realistic misinformation that people are likely to be confused about (Table 4), and are challenging to current state-of-the-art models (Section 4.1). Moreover, training a model on our data improves professional fact checking (Section 4.2).

QA to Verification Task Prior work has also used QA data to create entailment or fact verification benchmarks. Most make use of synthetic or annotated questions (Demszky et al., 2018; Jiang et al., 2020; Pan et al., 2021; Chen et al., 2021)[4]

[2]For this reason, prior work on professional fact checking assumes gold evidence document.

[3]Most claims fall into the `mixture` label, rather than `support` or `refute`.

[4]Annotated questions are simulated by crowdworkers given the evidence text and the answer, having largely different distributions from information-seeking questions (Lee et al., 2019; Gardner et al., 2019).

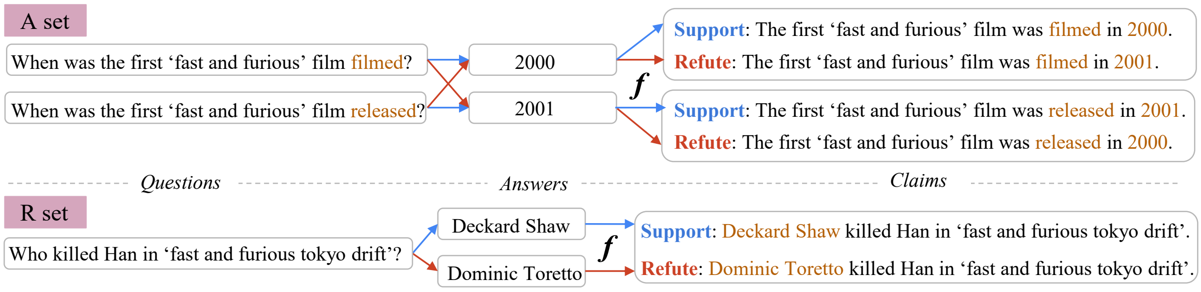

Figure 2: Overview of the data creation process. The data consits of two sets (A and R). For A, we use the disambiguated question-answer pairs and generate `support` and `refute` claims from matching pairs (*filmed–2000, released–2001*) and crossover pairs (*filmed–2001, released–2000*), respectively. For R, we use the reference answer (*Deckard Shaw*) and the incorrect prediction from DPR (*Dominic Toretto*) to generate `support` and `refute` claims, respectively. f is a T5 model that transforms question-answer pairs to claims (Section 3.1.3).

while we use questions posed by real users to reflect confusions that naturally occur while seeking information. Thorne et al. (2021) use information-seeking questions, by converting yes/no questions to `support`/`refute` claims, but at a small scale and with unambiguous questions. Instead, our work uses large-scale information-seeking questions (with no restriction in answers) to claims. We are also unique in using highly ambiguous QA pairs to obtain claims that are more challenging to verify and have significantly fewer lexical cues (quantitative comparisons in Section 3.3).

3 Data

3.1 Data Construction

We construct **FAVIQ—FAct Verification derived from Information-seeking Questions**, where the model is given a natural language claim and predicts `support` or `refute` with respect to the English Wikipedia. The key idea to construct the data is to gather a set of valid and invalid question-answer pairs (Section 3.1.2) from annotations of information-seeking questions and their ambiguities (Section 3.1.1), and then convert each question-answer pair (q, a) to a claim (Section 3.1.3). Figure 2 presents an overview of this process.

3.1.1 Data Sources

We use QA data from Natural Questions (NQ, Kwiatkowski et al. (2019)) and AmbigQA (Min et al., 2020). NQ is a large-scale dataset consisting of the English information-seeking questions mined from Google search queries. AmbigQA provides disambiguated question-answer pairs for NQ questions, thereby highlighting the ambiguity that is inherent in information-seeking questions. Given an ambiguous question, it provides a set of multiple distinct answers, each paired with a new disambiguated question that uniquely has that answer.

3.1.2 Composing Valid and Invalid QA Pairs

FAVIQ is constructed from ambiguous questions and their disambiguation (*A* set) and is further augmented by using unambiguous question-answer pairs (*R* set).

From ambiguous questions (*A* set) We use the data consisting of a set of $(q, \{q_1, a_1\}, \{q_2, a_2\})$, where q is an information seeking question that has a_1, a_2 as multiple distinct answers.[5] q_1 and q_2 are disambiguated questions for the answers a_1 and a_2, i.e., q_1 has a_1 as a valid answer and a_2 as an invalid answer. We use (q_1, a_1) and (q_2, a_2) as valid question-answer pairs, and (q_1, a_2) and (q_2, a_1) as invalid question-answer pairs.

This data is particularly well suited to fact checking because individual examples require identification of entities, events, or properties that are semantically close but distinct: the fact that a user asked an ambiguous question q without realizing the difference between (q_1, a_1) and (q_2, a_2) indicates that the distinction is non-trivial and is hard to notice without sufficient background knowledge about the topic of the question.

From regular questions (*R* set) We use the QA data consisting of a set of (q, a): an information-seeking question q and its answer a. We then obtain an invalid answer to q, denoted as a_{neg}, from an off-the-shelf QA model for which we use the model from Karpukhin et al. (2020)—DPR followed by a span extraction model. We choose a_{neg}

[5] If q has more than two distinct answers, we sample two. This is to construct a reasonable number of claims per q.

		Total	Support	Refute
Train	A	17,008	8,504	8,504
	R	140,977	70,131	70,846
Dev	A	4,260	2,130	2,130
	R	15,566	7,739	7,827
Test	A	4,688	2,344	2,344
	R	5,877	2,922	2,955

Table 1: FAVIQ statistics. *A* includes claims derived from ambiguous questions, while *R* includes claims from regular question-answer pairs.

Data	Size	Length of the claims			
		Avg	Q1	Q2	Q3
Professional claims					
SNOPES	16k	12.4	10	11	14
SCIFACT	1k	11.5	9	11	13
Crowdsourced claims					
FEVER	185k	9.3	7	9	11
FM2	13k	13.9	9	13	16.3
BOOLQ-FV	10k	8.7	8	8	9
FAVIQ	**188k**	**12.0**	**9**	**10.5**	**13.5**

Table 2: Statistics of a variety of fact verification datasets. *Avg* and *Q1–3* are the average and quantiles of the length of the claims based on whitespace tokenization on the validation data; for FAVIQ, we report the macro-average of the A set and the R set. dataname is as large as FEVER and has a distribution of claim lengths that is much closer to that of professional fact checking datasets (SNOPES and SCIFACT).

with heuristics to obtain hard negatives but not the false negative; details provided in Appendix A. We use (q, a) and (q, a_{neg}) as a valid and an invalid question-answer pair, respectively.

We can think of (q, a_{neg}) as a *hard* negative pair chosen adversarially from the QA model.[6] This data can be obtained on a much larger scale than the A set because annotating a single valid answer is easier than annotating disambiguations.

3.1.3 Transforming QA pairs to Claims

We transform question-answer pairs to claims by training a neural model which maps (q, a) to a claim that is `support` *if and only if* a is a valid answer to q, otherwise `refute`. We first manually convert 250 valid and invalid question-answer pairs obtained through Section 3.1.2 to claims. We then train a T5-3B model (Raffel et al., 2020), using 150 claims for training and 100 claims for validation. The model is additionally pretrained on data from Demszky et al. (2018), see Appendix A.

3.1.4 Obtaining silver evidence passages

We obtain silver evidence passages for FAVIQ by (1) taking the question that was the source of the claim during the data creation (either a user question from NQ or a disambiguated question from AmbigQA), (2) using it as a query for TF-IDF over the English Wikipedia, and (3) taking the top passage that contains the answer. Based on our manual verification on 100 random samples, the precision of the silver evidence passages is 70%. We provide silver evidence passages primarily for supporting training of the model, and do not explicitly evaluate passage prediction; more discussion in Appendix A. Future work may use human annotations on top of our silver evidence passages in order to further improve the quality, or evaluate passage prediction.

[6]It is possible that the R set contains bias derived from the use of DPR. We thus consider the R set as a source for data augmentation, while A provides the main data.

3.2 Data Validation

In order to evaluate the quality of claims and labels, three native English speakers were given 300 random samples from FAVIQ, and were asked to: (1) verify whether the claim is as natural as a human-written claim, with three possible ratings (perfect, minor issues but comprehensible, incomprehensible), and (2) predict the label of the claim (`support` or `refute`). Validators were allowed to use search engines, and were encouraged to use the English Wikipedia as a primary source.

Validators found 80.7% of the A set and 89.3% of the R set to be natural, and 0% to be incomprehensible. The rest have minor grammatical errors or typos, e.g., missing "the". In most cases the errors actually come from the original NQ questions which were human-authored, indicating that these grammatical errors and typos occur in real life. Lastly, validators achieved an accuracy of 95.0% (92.7% of A and 97.3% of R) when evaluated against gold labels in the data—this indicates high-quality of the data and high human performance. This accuracy level is slightly higher than that of FEVER (91.2%).

3.3 Data Analysis

Data statistics for FAVIQ are listed in Table 1. It has 188k claims in total, with balanced `support` and `refute` labels. We present quantitative and qualitative analyses showing that claims on FAVIQ contain much less lexical bias than other crowdsourced datasets and include misinformation that is realistic and harder to identify.

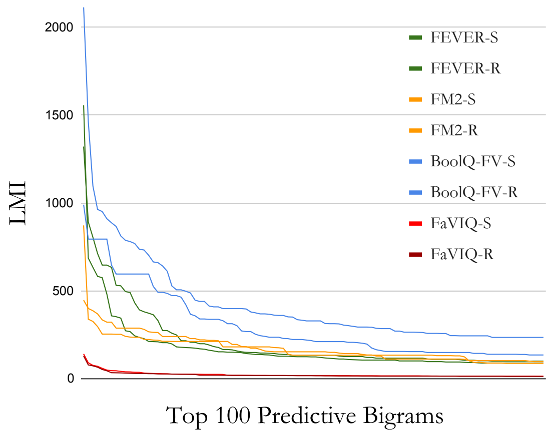

Figure 3: Plot of LMI scores of top 100 predictive bigrams for FEVER, FM2, BOOLQ-FV and FAVIQ (macro-averaged over the A set and the R set). `S` and `R` denotes `support` and `refute`, respectively. BOOLQ-FV indicates data from Thorne et al. (2021) that uses BOOLQ. LMI scores of FAVIQ are significantly lower than those of FEVER and FM2, indicating significantly less lexical overlap.

Comparison of size and claim length Table 2 compares statistics of a variety of fact verification datasets: SNOPES (Hanselowski et al., 2019), SCIFACT (Wadden et al., 2020), FEVER (Thorne et al., 2018a), FM2 (Eisenschlos et al., 2021), BOOLQ-FV (Thorne et al., 2021) and FAVIQ.

FAVIQ is as large-scale as FEVER, while its distributions of claim length is much closer to claims authored by professional fact checkers (SNOPES and SCIFACT). FM2 is smaller scale, due to difficulty in scaling multi-player games used for data construction, and has claims that are slightly longer than professional claims, likely because they are intentionally written to be difficult. BOOLQ-FV is smaller, likely due to relative difficulties in collecting naturally-occurring yes/no questions.

Lexical cues in claims We further analyze lexical cues in the claims on FEVER, FM2, BOOLQ-FV and FAVIQ by measuring local mutual information (LMI; Schuster et al. (2019); Eisenschlos et al. (2021)). LMI measures whether the given bigram correlates with a particular label. More specifically, LMI is defined as:

$$LMI(w,c) = P(w,c)\ log\frac{P(w,c)}{P(w)\cdot P(c)},$$

where w is a bigram, c is a label, and $P(\cdot)$ are estimated by counting (Schuster et al., 2019).

Dataset	Top Bigrams by LMI
FEVER-S	is a, a film, of the, is an, in the, in a
FEVER-R	is only, only a, incapable of, is not, was only, is incapable
A set-S	on the, was the, the date, date of, in episode, is what
A set-R	of the, the country, at the, the episode, started in, placed at
R set-S	out on, on october, on june, released on, be 18, on august
R set-R	out in, on september, was 2015, of the, is the, released in

Table 3: Top bigrams with the highest LMI for FEVER and FAVIQ. `S` and `R` denotes `support` and `refute` respectively. Highlighted bigrams indicate negative expressions, e.g., "only", "incapable" or "not".

The distributions of the LMI scores for the top-100 bigrams are shown in Figure 3. The LMI scores of FAVIQ are significantly lower than those of FEVER, FM2, and BOOLQ-FV, indicating that FAVIQ contains significantly less lexical bias.

Tables 3 shows the top six bigrams with the highest LMI scores for FEVER and FAVIQ. As highlighted, all of the top bigrams in `refute` claims of FEVER contain negative expressions, e.g., "is only", "incapable of", "did not". In contrast, the top bigrams from FAVIQ do not include obvious negations and mostly overlap across different labels, strongly suggesting the task has fewer lexical cues. Although there are still top bigrams from FAVIQ causing bias (e.g., related to time, such as 'on October'), their LMI values are significantly lower compared those from other datasets.

Qualitative analysis of the `refute` claims We also analyzed 30 randomly sampled `refute` claims from FAVIQ and FEVER respectively. We categorized the cause of misinformation as detailed in Appendix B, and show three most common categories for each dataset as a summary in Table 4.

On FAVIQ, 60% of the claims involve entities, events or properties that are semantically close, but still distinct. For example, they are specified with conjunctions (e.g., "was foreign minister" and "signed the treaty of versailles from germany"), or share key attributes (e.g., films with the same title). This means that relying on lexical overlap or partially understanding the evidence text would lead to incorrect predictions; one must read the full evidence text to realize that the claim is false. Furthermore, 16.7% involve events, e.g., from filing for bankruptcy for the first time to completely ceasing operations (Table 4). This requires full understanding of the underlying event and tracking of state changes (Das et al., 2019; Amini et al., 2020).

The same analysis on FEVER confirms the findings from Schuster et al. (2019); Eisenschlos et al.

Conjunctions (33.3%) *C*: Johannes bell was the foreign minister that signed the treaty of versailles from germany. / *E*: Johannes bell served as Minister of Colonial Affairs ... He was one of the two German representatives who signed the Treaty of Versailles. **Shared attributes (26.7%)** *C*: Judi bowker played andromeda in the 2012 remake of the 1981 film clash of the titans called wrath of the titans. *E*: Judi bowker ... Clash of the Titans (1981). **Procedural event (16.7%)** *C*: Mccrory's originally filed for bankruptcy on february 2002. / *E*: McCrory Stores ... by 1992 it filed for bankruptcy. ... In February 2002 the company ceased operation.
Negation (30.0%) *C*: Southpaw hasn't been released yet. *E*: Southpaw is an American sports drama film released on July 24, 2015. **Cannot find potential cause (20.0%)** *C*: Mutiny on the Bounty is Dutch. *E*: Mutiny on the Bounty is a 1962 American historical drama film. **Antonym (13.3%)** *C*: Athletics lost the world series in 1989. *E*: The 1989 World Series ... with the Athletics sweeping the Giants.

Table 4: Three most common categories based on 30 `refute` claims randomly sampled from the validation set, for FAVIQ (top) and FEVER (bottom) respectively. Full statistics and examples in Appendix B. *C* and *E* indicate the claim and evidence text, respectively. `Refute` claims in FAVIQ are more challenging, not containing explicit negations or antonyms.

(2021); many of claims contain explicit negations (30%) and antonyms (13%), with misinformation that is less likely to occur in the real world (20%).[7]

4 Experiments

We first evaluate state-of-the-art fact verification models on FAVIQ in order to establish baseline performance levels (Section 4.1). We then conduct experiments on professional fact-checking datasets to measure the improvements from training on FAVIQ (Section 4.2).

4.1 Baseline Experiments on FAVIQ

4.1.1 Models

We experiment with two settings: a zero-shot setup where models are trained on FEVER, and a standard setup where models are trained on FAVIQ. For FEVER, we use the KILT (Petroni et al., 2021) version following prior work; we randomly split the official validation set into equally sized validation and test sets, as the official test set is hidden.

[7]For instance, consider the claim "Mutiny on the Bounty is Dutch" in Table 4. There is no Dutch producer, director, writer, actors, or actress in the film—we were not able to find a potential reason that one would believe that the film is Dutch.

All models are based on BART (Lewis et al., 2020), a pretrained sequence-to-sequence model which we train to generate either `support` or `refute`. We describe three different variants which differ in their input, along with their accuracy on FEVER by our own experiments.

Claim only BART takes a claim as the only input. Although this is a trivial baseline, it achieves an accuracy of 79% on FEVER.

TF-IDF + BART takes a concatenation of a claim and k passages retrieved by TF-IDF from Chen et al. (2017). It achieves 87% on FEVER. We choose TF-IDF over other sparse retrieval methods like BM25 (Robertson and Zaragoza, 2009) because Petroni et al. (2021) report that TF-IDF outperforms BM25 on FEVER.

DPR + BART takes a concatenation of a claim and k passages retrieved by DPR (Karpukhin et al., 2020), a dual encoder based model. It is the state-of-the-art on FEVER based on Petroni et al. (2021) and Maillard et al. (2021), achieving an accuracy of 90%.

Implementation details We use the English Wikipedia from 08/01/2019 following KILT (Petroni et al., 2021). We take the plain text and lists provided by KILT and create a collection of passages where each passage has up to 100 tokens. This results in 26M passages. We set the number of input passages k to 3, following previous work (Petroni et al., 2021; Maillard et al., 2021). Baselines on FAVIQ are jointly trained on the A set and the R set.

Training DPR requires a positive and a negative passage—a passage that supports and does not support the verdict, respectively. We use the silver evidence passage associated with FAVIQ as a positive, and the top TF-IDF passage that is not the silver evidence passages as a negative. More training details are in Appendix C. Experiments are reproducible from `https://github.com/faviq/faviq/tree/main/codes`.

4.1.2 Results

Table 5 reports results on FAVIQ. The overall accuracy of the baselines is low, despite their high performance on FEVER. The zero-shot performance is barely better than random guessing, indicating that the model trained on FEVER is not able to generalize to our more challenging data. When the baselines are trained on FAVIQ, the best model

Model	Dev		Test	
	A	R	A	R
Training on FEVER *(zero-shot)*				
Clain only BART	51.6	51.0	51.9	51.1
TF-IDF + BART	55.8	58.5	54.4	57.2
DPR + BART	56.0	62.3	55.7	61.2
Training on FAVIQ				
Claim only BART	51.0	59.5	51.3	59.4
TF-IDF + BART	65.1	74.2	63.0	71.2
DPR + BART	**66.9**	**76.8**	**64.9**	**74.6**

Table 5: Fact verification accuracy on FAVIQ. DPR + BART achieves the best accuracy; however, there is overall significant room for improvement.

achieves an accuracy of 65% on the A set, indicating that existing state-of-the-art models do not solve our benchmark.[8]

Impact of retrieval The performance of the claim only baseline that does not use retrieval is almost random on FAVIQ, while achieving nearly 80% accuracy on FEVER. This result suggests significantly less bias in the claims, and the relative importance of using background knowledge to solve the task. When retrieval is used, DPR outperforms TF-IDF, consistent with the finding from Petroni et al. (2021).

A set vs. R set The performance of the models on the R set is consistently higher than that on the A set by a large margin, implying that claims based on ambiguity arisen from real users are more challenging to verify than claims generated from regular question-answer pairs. This indicates clearer contrast to prior work that converts regular QA data to declarative sentences (Demszky et al., 2018; Pan et al., 2021).

Error Analysis We randomly sample 50 error cases from DPR + BART on the A set of FAVIQ and categorize them, as shown in Table 6.

- *Retrieval error* is the most frequent type of errors. DPR typically retrieves a passage with the correct topic (e.g., about "Lie to Me") but that is missing more specific information (e.g., the end date). We think the claim having less lexical overlap with the evidence text leads to low recall@k of the retrieval model ($k = 3$).

[8]We additionally show and discuss the model trained on FAVIQ and tested on FEVER in Appendix D. They achieve non-trivial performance (67%) although being worse than FEVER-trained models that exploit bias in the data.

- 28% of error cases involve *events*. In particular, 14% involve procedural events, and 6% involve distinct events that share similar properties but differ in location or time frame.
- In 18% of error cases, retrieved evidence is *valid but not notably explicit*, which is naturally the case for the claims occurring in real life. FAVIQ has this property likely because it is derived from questions that are gathered independently from the evidence text, unlike prior work (Thorne et al., 2018a; Schuster et al., 2021; Eisenschlos et al., 2021) with claims written given the evidence text.
- 16% of the failure cases require *multi-hop* inference over the evidence. Claims in this category usually involve procedural events or compositions (e.g. "is Seth Curry's brother" and "played for Davidson in college"). This indicates that we can construct a substantial portion of claims requiring multi-hop inference without having to make data that artificially encourages such reasoning (Yang et al., 2018; Jiang et al., 2020).
- Finally, 10% of the errors were made due to a subtle mismatch in *properties*, e.g., in the example in Figure 6, the model makes a decision based on "required minimum number" rather than "exact number" of a particular brand.

4.2 Professional Fact Checking Experiments

We use two professional fact-checking datasets.

SNOPES (Hanselowski et al., 2019) consists of 6,422 claims, authored and labeled by professional fact-checkers, gathered from the Snopes website.[9] We use the official data split.

SCIFACT (Wadden et al., 2020) consists of 1,109 claims based on scientific papers, annotated by domain experts. As the official test set is hidden, we use the official validation set as the test set, and separate the subset of the training data as the validation set to be an equal size as the test set.

For both datasets, we merge `not enough info` (`NEI`) to `refute`, following prior work that converts the 3-way classification to the 2-way classification (Wang et al., 2019; Sathe et al., 2020; Petroni et al., 2021).

4.2.1 Models

As in Section 4, all models are based on BART which is given a concatenation of the claim and

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

Category	%	Example
Retrieval error	38	*C*: The american show lie to me ended on january 31, 2011. (Support; Refute) *E*: Lie to Me ... The second season premiered on September 28, 2009 ... The third season, which had its premiere moved forward to October 4, 2010.
Events	28	*C*: The bellagio in las vegas opened on may, 1996. (Refute; Support) *E*: Construction on the Bellagio began in May 1996. ... Bellagio opened on October 15, 1998.
Evidence not explicit	18	*C*: The official order to start building the great wall of china was in 221 bc. (Support; Refute) *E*: The Great Wall of China had been built since the Qin dynasty (221–207 BC).
Multi-hop	16	*C*: Seth curry's brother played for davidson in college. (Support; Refute) *E*: Stephen Curry (...) older brother of current NBA player Seth ... He ultimately chose to attend Davidson College, who had aggressively recruited him from the tenth grade.
Properties	10	*C*: The number of cigarettes in a pack of 'export as' brand packs in the usa is 20. (Refute; Support) *E*: In the United States, the quantity of cigarettes in a pack must be at least 20. Certain brands, such as Export As, come in packs of 25.
Annotation error	4	*C*: The place winston moved to in still game is finport. (Refute; Support)

Table 6: Error analysis on 50 samples of the A set of FAVIQ validation data. *C* and *E* indicate the claim and retrieved evidence passages from DPR, respectively. Gold and blue indicate gold label and prediction by the model, respectively. The total exceeds 100% as one example may fall into multiple categories.

the evidence text and is trained to generate either `support` or `refute`. For SNOPES, the evidence text is given in the original data. For SCIFACT, the evidence text is retrieved by TF-IDF over the corpus of abstracts from scientific papers, provided in the original data. We use TF-IDF over DPR because we found DPR works poorly when the training data is very small.

We consider two settings. In the first setting, we assume the target training data is unavailable and compare the model trained on FEVER and FAVIQ in a zero-shot setup. In the second setting, we allow training on the target data and compare the model trained on the target data only and the model with the transfer learning—pretrained on either FEVER or FAVIQ and finetuned on the target data.

To explore models pretrained on `NEI` labels, we add a baseline that is trained on a union of the KILT version of FEVER and `NEI` data from the original FEVER from Thorne et al. (2018a). For FAVIQ, we also conduct an ablation that includes the R set only or the A set only.

Implementation details When using TF-IDF for SCIFACT, we use a sentence as a retrieval unit, and retrieve the top 10 sentences, which average length approximates that of 3 passages from Wikipedia. When using the model trained on either FEVER or FAVIQ, we use DPR + BART by default, which gives the best result in Section 4.1. As an exception, we use TF-IDF + BART on SCIFACT for a more direct comparison with the model trained on the target data only that uses TF-IDF.

When the models trained on FEVER or FAVIQ are used for professional fact checking, we find models are poorly calibrated, likely due to a domain shift, as also observed by Kamath et al. (2020) and Desai and Durrett (2020). We therefore use a simplified version of Platt scaling, a post-hoc calibration method (Platt et al., 1999; Guo et al., 2017; Zhao et al., 2021). Given normalized probabilities of `support` and `refute`, denoted as p_{s} and p_{r}, modified probabilities p'_{s} and p'_{r} are obtained via:

$$\begin{bmatrix} p'_{\mathrm{s}} \\ p'_{\mathrm{r}} \end{bmatrix} = \mathrm{Softmax}\left(\begin{bmatrix} p_{\mathrm{s}} + \gamma \\ p_{\mathrm{r}} \end{bmatrix}\right),$$

where $-1 < \gamma < 1$ is a hyperparameter tuned on the validation set.

4.2.2 Results

Table 7 reports accuracy on professional fact-checking datasets, SNOPES and SCIFACT.

Impact of transfer learning We find that transfer learning is effective—pretraining on large, crowdsourced datasets (either FEVER or FAVIQ) and finetuning on the target datasets always helps. Improvements are especially significant on SCIFACT, likely because its data size is smaller.

Using the target data is still important—models finetuned on the target data outperform zero-shot models by up to 20%. This indicates that crowdsourced data cannot completely replace professional fact checking data, but transfer learning from crowdsourced data leads to significantly better professional fact checking performance.

Training	SNOPES	SCIFACT
Majority only	60.1	58.7
No target data (zero-shot)		
FEVER	61.6	70.0
FEVER w/ NEI	63.4	73.0
FAVIQ	**68.2**	**74.7**
FAVIQ w/o A set	63.1	73.3
FAVIQ w/o R set	66.3	68.7
Target data available		
Target only	80.6	62.0
FEVER→target	80.6	76.7
FEVER w/ NEI→target	81.6	77.0
FAVIQ→target	**82.2**	**79.3**
FAVIQ w/o A set→target	81.6	78.3
FAVIQ w/o R set→target	81.7	76.7

Table 7: Accuracy on the test set of professional fact-checking datasets. Training on FAVIQ significantly improves the accuracy on SNOPES and SCIFACT, both in the zero-shot setting and in the transfer learning setting.

FAVIQ vs. FEVER Models that are trained on FAVIQ consistently outperform models trained on FEVER, both with and without the target data, by up to 4.8% absolute. This demonstrates that FAVIQ is a more effective resource than FEVER for professional fact-checking.

The model on FEVER is more competitive when NEI data is included, by up to 3% absolute. While the models on FAVIQ outperform models on FEVER even without NEI data, future work can possibly create NEI data in FAVIQ for further improvement.

Impact of the A set in FAVIQ The performance of the models that use FAVIQ substantially degrades when the A set is excluded. Moreover, models trained on the A set (without R set) perform moderately well despite its small scale, e.g., on SNOPES, achieving the second best performance following the model trained on the full FAVIQ. This demonstrates the importance of the A set created based on ambiguity in questions.

SNOPES benefits more from the A set than the R set, while SCIFACT benefits more from the R set than the A set. This is likely because SCIFACT is much smaller-scale (1k claims) and thus benefits more from the larger data like the R set. This suggests that having both the R set and the A set is important for performance.

5 Conclusion & Future Work

We introduced FAVIQ, a new fact verification dataset derived from ambiguous information-seeking questions. We incorporate facts that real users were unaware of when posing the question, leading to false claims that are more realistic and challenging to identify without fully understanding the context. Our extensive analysis shows that our data contains significantly less lexical bias than previous fact checking datasets, and include `refute` claims that are challenging and realistic. Our experiments showed that the state-of-the-art models are far from solving FAVIQ, and models trained on FAVIQ lead to improvements in professional fact checking. Altogether, we believe FAVIQ will serve as a challenging benchmark as well as support future progress in professional fact-checking.

We suggest future work to improve the FAVIQ model with respect to our analysis of the model prediction in Section 4.1.2, such as improving retrieval, modeling multi-hop inference, and better distinctions between entities, events and properties. Moreover, future work may investigate using other aspects of information-seeking questions that reflect facts that users are unaware of or easily confused with. For example, one can incorporate false presuppositions in questions that arise when users have limited background knowledge (Kim et al., 2021). As another example, one can explore generating NEI claims by leveraging unanswerable information-seeking questions. Furthermore, FAVIQ can potentially be a challenging benchmark for the claim correction, a task recently studied by Thorne and Vlachos (2021) that requires a model to correct the `refute` claims.

Acknowledgements

We thank Dave Wadden, James Thorn and Jinhyuk Lee for discussion and feedback on the paper. We thank James Lee, Skyler Hallinan and Sourojit Ghosh for their help in data validation. This work was supported by NSF IIS-2044660, ONR N00014-18-1-2826, an Allen Distinguished Investigator Award, a Sloan Fellowship, National Research Foundation of Korea (NRF-2020R1A2C3010638) and the Ministry of Science and ICT, Korea, under the ICT Creative Consilience program (IITP-2022-2020-0-01819).

References

Aida Amini, Antoine Bosselut, Bhavana Dalvi, Yejin Choi, and Hannaneh Hajishirzi. 2020. Procedural reading comprehension with attribute-aware context flow. In *AKBC*.

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 *EMNLP*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL*.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can nli models verify qa systems' predictions? In *EMNLP: Findings*.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational journalism: A call to arms to database researchers. In *CIDR*.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. 2019. Building dynamic knowledge graphs from text using machine reading comprehension. In *ICLR*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *EMNLP*.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from wikipedia gamification. In *NAACL*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *NAACL*.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; when is it useful? *arXiv*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *CoNLL*.

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 *EMNLP: Findings*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *ACL*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *ACL*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a benchmark for question answering research. *TACL*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oğuz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-task retrieval for knowledge-intensive tasks. In *ACL*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *EMNLP*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *ACL*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktäschel, et al. 2021. KILT: a benchmark for knowledge intensive language tasks. In *NAACL*.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*.

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*.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from wikipedia. In *LREC*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. In *NAACL*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *EMNLP*.

James Thorne, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based verification for real world information needs. *arXiv preprint arXiv:2104.00640*.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against fact extraction and verification. *arXiv*.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *ACL*.

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

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The Fact Extraction and VERification (FEVER) shared task. In *FEVER: Workshop*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *ACL: Workshop (LTSCC)*.

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 *EMNLP*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *NeurIPS*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *ICML*.

A Details in Data Construction

Details of obtaining a_{neg} We obtain an invalid answer to the question, denoted as a_{neg}, using an off-the-shelf QA model, for which we use DPR followed by a span extractor (Karpukhin et al., 2020).

The most naive way to obtain a_{neg} is to take the highest scored prediction that is not equal to a. We however found such prediction is likely to be a valid answer to q, either because it is semantically the same as a, or because the ambiguity in the question leads to multiple distinct valid answers. We therefore use two heuristics that we find greatly reduce such false negatives. First, instead of taking the top incorrect prediction, we obtain the top k predictions $p_1...p_k$ from the model and randomly sample one from $\{p_1...p_k\} \setminus \{a\}$. We use $k = 50$.

Although this is not a fundamental solution to remove false negatives, it significantly alleviates the problem, drastically dropping the portion of false negatives from 14% to 2% based on our manual verification on 50 random samples. Second, we train a neural model that is given a pair of the text and classifies whether they are semantically equivalent or not. This model is based on T5-large, trained and validated respectively on 150 and 100 pairs of (a, p_i) $(i = 1...k)$ which we manually label. We then exclude the predictions in $\{p_1...p_k\}$ which are classified as semantically equivalent to a by the classifier.

QA-to-claim converter We use a pretrained sequence-to-sequence model trained on a small number of our own annotations. We first manually write 250 claims given valid or invalid question-answer pairs. We then train a T5-3B model (Raffel et al., 2020), using 150 claims for training and 100 claims for validation. Each question-answer pair is fed into T5 with special tokens `question:` and `answer:`, respectively.

When training, we evaluate on the validation data every epoch and stop training when the validation accuracy does not increase for ten epochs. The accuracy is measured by the exact match score of the generated and the reference text after normalization, which we found to correlate well with the quality of the generated claims. The final model we train achieves 83% on the validation data. At inference time, we filter claims that do not contain the answer string, which may happen when the question is overly specific.

Why don't we evaluate evidence prediction? Unlike FEVER (Thorne et al., 2018a), which includes evidence prediction as part of the task, our paper does not report the evidence prediction performance and mainly reports the classification accuracy. There are three reasons for this change:

1. As claims on FAVIQ were written independent from any reference text, gold evidence text must be gathered through a separate process, which greatly increases the cost. This is different from other annotated fact checking datasets where a crowdworker wrote a claim based on the reference text and therefore the same reference text can be considered as gold evidence.
2. Finding gold evidence text is an inherently incomplete process; no human can get close to, or even measure the upperbound. Therefore, even after exhaustive human annotation, evaluation against annotated evidence leads to significant amount of false negatives. For example, when manually evaluating the top negatives of TF-IDF on 50 random samples from FEVER, 42% are false negatives.
3. Including evidence prediction as part of evaluation significantly restricts the approach models can take. For instance, one may choose not to use the text corpus provided in the dataset (e.g., Wikipedia), and decide to use other sources such as structured data (e.g. knowledge bases) or implicit knowledge stored in large neural models.

Nonetheless, as described in Section 3.1.4, we still provide the silver evidence passages which is useful to train a model, e.g., DPR, and supports future work to evaluate the evidence prediction accuracy.

B Analysis of `refute` Claims

We randomly sample 30 `refute` claims from FAVIQ and FEVER, respectively, and categorize the cause of the misinformation, as shown in Table 8. See Section 3.3 for discussion.

C Details of Experiments

DPR training for FEVER As FEVER provides the annotated evidence passage, we use it as a positive training example. We obtain a negative by querying the claim to TF-IDF and taking the passage that is not the positive passage and has the second highest score. We initially considered using the negative with the highest score, but found that

Category	% FAVIQ	% FEVER	Example
Negation	0	30.0	*C*: Southpaw hasn't been released yet. (from FEVER) *E*: Southpaw is a 2015 American sports drama film ... released on July 24, 2015.
Antonym	3.3	13.3	*C*: Athletics lost the world series in 1989. *E*: The 1989 World Series was ... with the Athletics sweeping the Giants in four games.
Requires reading across conjunctions	33.3	6.6	*C*: Johannes bell was the foreign minister that signed the treaty of versailles from germany. / *E*: Johannes bell served as Minister of Colonial Affairs ... He was one of the two German representatives who signed the Treaty of Versailles.
Shared attributes	26.7	6.6	*C*: Judi bowker played andromeda in the 2012 remake of the 1981 film clash of the titans called wrath of the titans. / *E*: Judi bowker ... Clash of the Titans (1981).
Procedural event	16.7	0	*C*: Mccrory's originally filed for bankruptcy on february 2002. / *E*: McCrory Stores ... by 1992 it filed for bankruptcy. ... In February 2002 the company ceased operation.
Incorrect type of properties	10.0	3.3	*C*: Tyler, the Creator is the name of the song at the end of who dat boy. *E*: "Who Dat Boy" is a song by American rapper Tyler, the Creator.
Cannot find potential cause	0	20.0	*C*: Mutiny on the Bounty is Dutch. (from FEVER) *E*: Mutiny on the Bounty is a 1962 American Technicolor epic historical drama film.
Annotation error	10.0	20.0	*C*: Pasek and paul were the individuals that wrote the lyrics to the greatest showman.

Table 8: Categorization of 30 `refute` claims on FAVIQ and FEVER, randomly sampled from the validation set. *C* and *E* indicate the claim and evidence text, respectively. Examples are from FAVIQ unless otherwise specified.

Model	Dev	Test
Clain only BART	47.2	48.3
TF-IDF + BART	67.8	66.6
DPR + BART	67.2	66.5

Table 9: Fact verification accuracy on FEVER of different models when trained on FAVIQ.

many of them (37%) are false negatives based on our manual evaluation of 30 random samples. This is likely due to incomprehensive evidence annotation as discussed in Appendix A. We find using the negative with the second highest instead decreases the portion of false negatives from 37% to 13%.

Other details Our implementations are based on PyTorch[10] (Paszke et al., 2019) and Huggingface Transformers[11] (Wolf et al., 2020).

When training a BART-based model, we map `support` and `refute` labels to the words 'true' and 'false' respectively so that each label is mapped to a single token. This choice was made against mapping to 'support' and 'refute'because the BART tokenizer maps 'refute' into two tokens, making it difficult to compare probabilities of `support` and `refute`.

By default, we use a batch size of 32, a maximum sequence length of 1024, and 500 warmup steps using eight 32GB GPUs. For SCIFACT, we use a batch size of 8 and no warmup steps using four 32G GPUs. We tune the learning rate in between {7e-6, 8e-6, 9e-6, 1e-5} on the validation data.

D Additional Experiments

Table 9 reports the model performance when trained on FAVIQ and tested on FEVER. The best-performing model achieves non-trivial performance (67%). However, their overall performance is not as good as model performance when trained on FEVER, likely because the models do not exploit the bias in the FEVER dataset. Nonetheless, we underweight the test performance on FEVER due to known bias in the data.

[10] `https://pytorch.org/`

[11] `https://github.com/huggingface/transformers`