How Robust are Fact Checking Systems on Colloquial Claims?

Byeongchang Kim* **Hyunwoo Kim*** **Seokhee Hong** **Gunhee Kim**
Seoul National University, Seoul, Korea
{byeongchang.kim,hyunw.kim,seokhee.hong}@vl.snu.ac.kr, gunhee@snu.ac.kr
https://vl.snu.ac.kr/projects/colloquial-claims

Abstract

Knowledge is now starting to power neural dialogue agents. At the same time, the risk of misinformation and disinformation from dialogue agents also rises. Verifying the veracity of information from formal sources are widely studied in computational fact checking. In this work, we ask: *How robust are fact checking systems on claims in colloquial style?* We aim to open up new discussions in the intersection of fact verification and dialogue safety. In order to investigate how fact checking systems behave on colloquial claims, we transfer the styles of claims from FEVER (Thorne et al., 2018) into colloquialism. We find that existing fact checking systems that perform well on claims in formal style significantly degenerate on colloquial claims with the same semantics. Especially, we show that document retrieval is the weakest spot in the system even vulnerable to filler words, such as "*yeah*" and "*you know*". The document recall of WikiAPI retriever (Hanselowski et al., 2018) which is 90.0% on FEVER, drops to 72.2% on the colloquial claims. We compare the characteristics of colloquial claims to those of claims in formal style, and demonstrate the challenging issues in them.

1 Introduction

Recently, knowledge has been starting to power neural dialogue agents (Moghe et al., 2018; Zhou et al., 2018b; Ghazvininejad et al., 2018; Qin et al., 2019; Gopalakrishnan et al., 2019), being equipped with Wikipedia (Dinan et al., 2019b), news (Gopalakrishnan et al., 2019), domain specific knowledge-base (Eric and Manning, 2017), and commonsense (Zhou et al., 2018a; Young et al., 2018; Wu et al., 2020). However, the use of knowledge inevitably put dialogue agents in new jeopardy. For example, recent workshop on safety for conversational AI (Dinan et al., 2020b) introduced an example of such risk: Bickmore et al. (2018) asked participants to query conversational agents for advice in situations where medical information is needed. Then, internist and pharmacist judged the actions that the participants would take based on the advice. Assessments revealed that agents often deliver incorrect medical information that may cause lethal consequences.

A bigger threat may be the abuse of dialogue agents to deliberately distribute disinformation. What would happen if knowledge-powered agents are tweaked to massively generate false claims on online communities? The impact of such fake news can be critical as they quickly spread through social media (Shu et al., 2017). The chatbot Tay's shut down due to malicious attempts show the imminent danger of abuse (Wolf et al., 2017).

Verifying the integrity of a given piece of information has been studied in the field of computational fact checking. Thorne et al. (2018) introduce an annotated dataset FEVER for fact checking based on Wikipedia. Augenstein et al. (2019) collect claims on fact checking websites and release the MultiFC dataset. Jiang et al. (2020) collect a dataset requiring many-hop evidence extraction from Wikipedia. Wadden et al. (2020) collect a dataset of scientific claims to be verified.

Most claims of existing datasets are taken from formal texts, such as news, academic papers, and Wikipedia. These claims tend to be concise and structured: "*Beautiful was number two on the Billboard Hot 100 in 2003*". On the other hand, claims or information that we encounter in dialogues are more unstructured and informal: "*The song Beautiful is great! It even reached number two on the Hot 100 in 2003, you know?*". For improving the applicability of fact checking systems, they must also be robust for verifying the claims in dialogues.

Unfortunately, threats regarding misinformation and disinformation from dialogue agents remain understudied. Research on dialogue safety mainly

*Equal contribution

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1535–1548
June 6–11, 2021.

has focused on making dialogue agents robust to adversarial attacks (Dinan et al., 2019a), and preventing dialogue agents from generating offensive or biased responses (Henderson et al., 2018; Sap et al., 2019; Xu et al., 2020).

In this work, we aim to investigate how fact checking systems behave when verifying claims in dialogue style, rather than claims from news outlets, scientific articles, or Wikipedia. Colloquial claims are different in several aspects compared to claims from formal sources. (i) They tend to also include filler words, casual comments, or personal feelings which do not require verification. (ii) Since claims in colloquial language are less precise than formal claims, correctly using the context in claims becomes important to disambiguate them. We demonstrate that these features make existing fact checking systems have difficulties in verifying colloquial claims. We use English datasets for the investigation in this work. Our major contributions of this work can be outlined as follows:

(1) We open up new discussions in the intersection of fact verification and dialogue safety; how to verify claims in colloquial language, compared to previous works that solely focus on the claims in formal style (*e.g.* news, academic papers, Wikipedia).

(2) For this study, we curate colloquial claims by transferring the styles of claims in existing fact checking dataset of FEVER (Thorne et al., 2018). For style transfer, we finetune a pretrained dialogue model with a knowledge-grounded dialogue dataset and apply additional filtering to compensate for the quality of output.

(3) We show that the existing fact checking systems that perform well on claims in formal style significantly degenerate on colloquial ones with the same semantics. We analyze the performance drop and show document retrieval is the weakest spot in the system.

(4) We identify the challenging characteristics of colloquial claims; (i) they often involve expressions that are not verifiable (*e.g.* filler words or personal feeling) and (ii) they include ambiguity inside the claim that necessities better understanding of the context. We release the code and the curated colloquial claims set.

FEVER (Thorne et al., 2018)
Claim: The iPhone 4 is a dial telephone. **Wikipedia Document**: [iPhone] **Evidence Sentence**: The iPhone 4 is a smartphone that was designed and marketed by Apple Inc.. **Label**: REFUTED
Wizard of Wikipedia (Dinan et al., 2019b)
Topic: [The Hershey Company] **Wikipedia Knowledge**: Headquarters are in Hershey, Pennsylvania, which is home to Hershey's Chocolate World. **Apprentice** (*i.e.* dialogue context): I love chocolate, my favorite is Hershey. What's yours? **Wizard**: I love Hershey too! Do you know that Hershey's HQ is actually located in Hershey, Pennsylvania?

Table 1: Example of FEVER and Wizard of Wikipedia.

2 Background

2.1 Fact Checking Pipeline

FEVER (Thorne et al., 2018) is a fact checking benchmark dataset based on Wikipedia. Its fact checking pipeline has become one of the standard followed by many (Hanselowski et al., 2018; Nie et al., 2019; Zhou et al., 2019; Liu et al., 2020; Zhong et al., 2020; Jiang et al., 2020). The pipeline comprises three stages: document retrieval, evidence selection, and claim verification. For a given claim to be verified, the system first retrieves the related documents from the pool. Next, among the returned documents, the system selects the most suitable sentences for evidence. Finally, based on the evidence sentences the system classifies the claim's veracity with three classes: SUPPORTED, REFUTED (contradicted by the evidence), and NOTENOUGHINFO (cannot be determined by the evidence). An example from the FEVER is shown in Table 1.

2.2 Wizard of Wikipedia

The Wizard of Wikipedia (WoW) (Dinan et al., 2019b) may be the closest dialogue dataset to existing fact checking datasets. It is a knowledge-based open-domain dialogue dataset involving two speakers discussing on a given topic. An example is presented in Table 1. One speaker (referred as apprentice) is eager to learn about the topic, while the other speaker (the wizard) delivers knowledge-grounded responses based on both dialogue context and Wikipedia documents for the topic. In this dataset, the gold "knowledge sentence" from Wikipedia is provided for each wizard's response. Hence, we can regard the gold knowledge sentence as the *evidence* for the Wizard's response.

However, WoW only provides pairs of (knowledge sentence, grounded response), hence those responses are all SUPPORTED by Wikipedia. There are no REFUTED or NOTENOUGHINFO responses in the dataset. Such limitation make it difficult to directly adopt WoW as a fact checking dialogue dataset. Nonetheless, its knowledge-grounding property makes it a useful resource for training dialogue models to generate colloquial utterances grounded on claims.

3 Transferring to Colloquial Claims

Our goal is to curate colloquial claims by transferring the style of each claim sentence in the FEVER dataset[1] into colloquial style. We first finetune a dialog model with the WoW dataset so that it learns to transfer knowledge sentences from Wikipedia into conversational utterances (section 3.1). We then apply the finetuned model to transfer each claim in FEVER (sourced from Wikipedia) into colloquial style, and perform filtering process to warrant the integrity of this style transfer (section 3.2). Figure 1 overviews the whole pipeline of style transfer.

3.1 Finetuning a Dialogue Model

We first finetune BART-large (Lewis et al., 2020) to generate the wizard's response given only the corresponding knowledge sentence from WoW, without the dialogue context. Take the example in Table 1, when the knowledge sentence is given as "*Hershey's headquarters are in Hershey, Pennsylvania*", BART is finetuned to generate the wizard's response "*I love Hershey too! Do you know that Hershey's HQ is actually in Hershey?*". We exclude the dialogue context during fine-tuning in order to enforce the dialogue model to exclusively focus on knowledge contents. The finetuned BART shows a low perplexity of 10.51 on WoW's validation set. This indicates that BART can generate information-grounded utterances when given knowledge sentences.

Then, we apply the finetuned BART to transfer each claim in FEVER to a colloquial one. Our expectation is that since claims in FEVER are based on Wikipedia too and similar to knowledge sentences in WoW in many aspects, the finetuned model may be able to produce utterances while preserving the semantics of claims from FEVER.

[1]We verified that FEVER is released under a Creative Commons (CC BY-SA 3.0) license.

However, naively using the generated claims as is has several issues, including (i) copy-and-paste, (ii) pronoun overwrite, (iii) semantic discrepancy, and (iv) lack of colloquialism. We carefully mitigate these issues through a filtering pipeline.

3.2 Oversampling and Filtering

We first oversample n colloquial candidates $Q_i = \{q_{i,j}\}_{j=1}^{468}$ per claim c_i in FEVER, using BART through Nucleus Sampling (Holtzman et al., 2020) ($p = 0.95$).

Preventing Copy-Paste. We observe the dialogue model sometimes simply copies the input claim as output. Since copy-pasted candidates are not colloquial, we remove the ones whose F1 scores are higher than 0.9, in respect to the original claim.

Preserving Named Entities. Utterances in dialogues tend to refer entities with pronouns rather than their original word. As a result, we observe that dialogue models also convert entities in claims to pronouns. For example, given the input claim "*Tetris has sold millions of physical copies*", BART outputs "*Yeah it's fun even today, no wonder it sold millions of physical copies*". Since there are no previous contexts for claims in FEVER, it is not possible to recognize that pronoun "*it*" is referring to "*tetris*".

In order to preserve the entities, we leverage the named entity recognition (NER) module from Stanza (Qi et al., 2020), which shows 88.8 F1-score on OntoNotes (Weischedel et al., 2013) test set. We extract a set of named entities $\mathcal{E}_i^c$ from claim c_i, and compare it with the named entity set $\mathcal{E}_{i,j}^q$ of each $q_{i,j}$ in Q_i. We remove candidates with less than two matching named entities. For claims with single named entity, we remove candidates having no named entities.

Preserving Semantic Equivalence. It is well known that neural dialogue models lack consistency (Li et al., 2016) and can hallucinate irrelevant content (Roller et al., 2020). As a result, there can be semantic difference between the original FEVER claim and the generated one.

To preserve the original semantics, we leverage natural language inference (NLI), which is a task of determining whether a hypothesis sentence can be inferred from the given premise sentence. The hypothesis sentence is classified into three categories: ENTAILMENT (true), CONTRADICTION (false), and NEUTRAL (undetermined). A sound colloquial claim should be entailed by the original

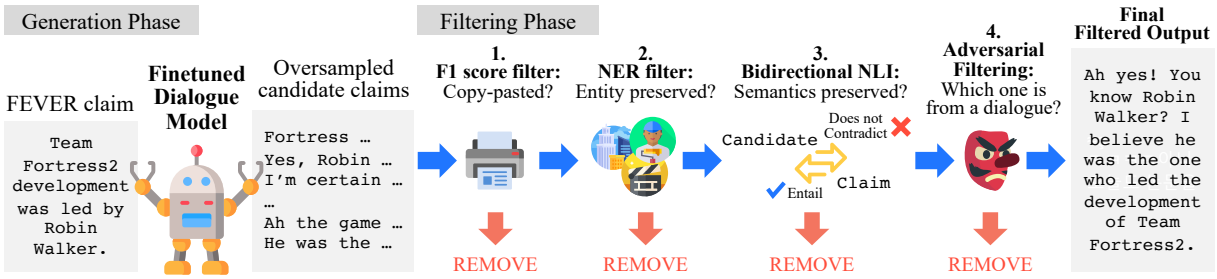

Figure 1: Illustration of the transfer pipeline for our Colloquial Claims.

	F1-score	NER	NLI	AF
Avg. Cumulative Survival Rate (%)	96.2	46.4	6.3	top-k

Table 2: The average cumulative survival rate of candidates after each filtering. We apply filters in the order of F1, NER, NLI, and AF.

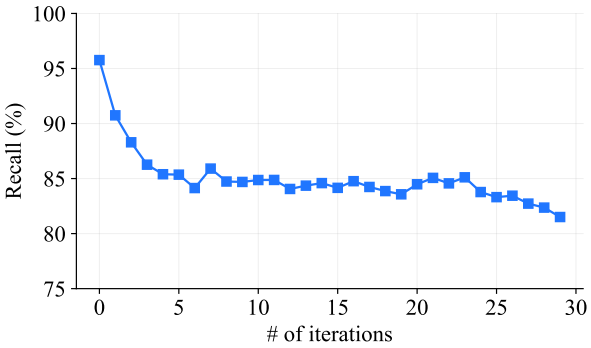

Figure 2: The recall for AFLITE linear classifiers for our Colloquial Claims.

claim and it also must not contradict the original. Suppose the original claim is "*Apple Inc. designed and manufactured iPhone 4*" and the generated claim is "*I heard Apple is also famous for designing the iMac computer*". This claim is removed because "designing iMac" cannot be inferred from the fact "Apple manufactured iPhone 4".

We conduct bidirectional NLI between the original claim and the generated one using RoBERTa (Liu et al., 2019) trained on MNLI (Williams et al., 2018). The RoBERTa model shows 90.59% accuracy on MNLI validation set. For each candidate $q_{i,j}$, we conduct NLI$(c_i, q_{i,j})$ and NLI$(q_{i,j}, c_i)$ with the original claim c_i. We only preserve the candidates that result in ENTAILMENT for the former and do not result in CONTRADICTION for the latter.

Ensuring Colloquialism. Although the candidates are generated by a dialogue model, they may still resemble the style of the original claims, rather than colloquial style. To ensure colloquialism, we select the top-k candidate claims which are most difficult to discriminate from responses in Wizard of Wikipedia (WoW) (Dinan et al., 2019b), through an iterative adversarial filtering method AFLITE (Sakaguchi et al., 2020; Bras et al., 2020). We first embed the candidates with RoBERTa and train an ensemble of binary linear classifiers to determine each candidate whether it is from WoW or our colloquial claims. We eliminate candidates that are easily classified as our colloquial claims after each iteration. We continue the iteration until k candidates remain in each Q_i. We set $k = 3$. Since only candidates that are hard to discriminate from WoW responses survive, they resemble the styles of dialogue utterances. We defer the detailed algorithm for adversarial filtering to Appendix.

Filtering Statistics. Table 2 shows the average survival rate of candidates after each filtering step. We observe that the NER and NLI filter effectively remove large amounts of candidates. On average, 29 out of 486 candidates survive after the NLI filtering stage. Then, adversarial filtering is used for selecting k candidates among the remainders.

Figure 2 shows the recall for our colloquial claims by the binary classifiers used in AFLITE. As only indistinguishable candidate claims from the WoW responses survive, the recall drops after each iteration. We also compare the qualitative traits of candidates before and after the filtering in Section 4.2.

3.3 Manual Quality Check on Test set

Finally, we manually check all SUPPORTED and REFUTED instances in the test set of our Colloquial Claims dataset. Three human annotators choose the best suitable claim for each colloquial claim set ($|Q_i| \leq k$) for the given label and evidence. If there are no suitable claim in the set, we recover the set before top-k selection. As a last resort, we let annotators rewrite the colloquial claim when no eligible candidate exists. The proportion that requires manual rewriting is less than 1% of 5,615

	#Claims			#Words/Claim
	Train	Valid	Test	
FEVER	145.4K	10K	10K	8.2
Colloquial Claims	410.0k	28.9K	8.4K	11.1

Table 3: Statistics of the Colloquial Claims compared to FEVER (Thorne et al., 2018).

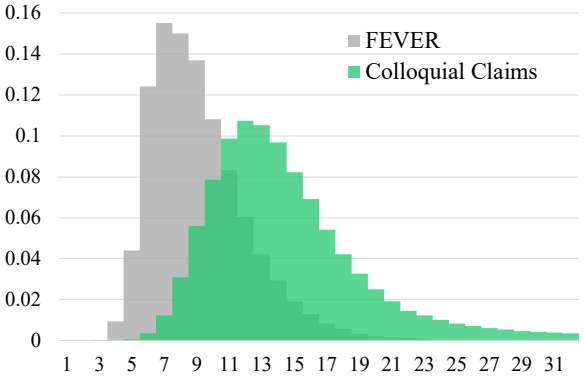

Figure 3: Comparison of the claim sentence length distributions between FEVER and our Colloquial Claims.

instances.

4 Properties of Colloquial Claims

4.1 Quantitative Comparison

We first discuss the characteristics of our Colloquial Claims with quantitative analysis, compared to FEVER (Thorne et al., 2018) and Wizard of Wikipedia (WoW) (Dinan et al., 2019b).

Diverse Claims. We provide basic statistics of our Colloquial Claims in Table 3. In FEVER, only a single claim exists per evidence set, whereas our Colloquial Claims provide up to three claims. As a result, the number of data instances of our dataset is larger than FEVER.

Due to the wordy nature of colloquial language, the our transferred claims are longer and more diverse in length than those in FEVER. Figure 3 plots the density of the claim sentence lengths of FEVER and our dataset.

Colloquial Style. The claims in our Colloquial Claims have similar styles to the utterances in dialogues. Following Yang et al. (2020), we gauge the style of sentences by measuring the perplexity with a pretrained DialoGPT (Zhang et al., 2019). The perplexity of the sentence becomes high if its style is far from a dialogue. Table 4 compares the perplexity of responses from WoW, claims from FEVER and our Colloquial Claims. The perplexity of claims in FEVER is high, whereas our Colloquial Claims have closer perplexity to WoW.

	WoW	FEVER	Colloquial Claims
DialoGPT Perplexity	471.9	1381.5	575.8

Table 4: The perplexity measured by DialoGPT (Zhang et al., 2019) for the responses in Wizard of Wikipedia (WoW) (Dinan et al., 2019b), claims in FEVER (Thorne et al., 2018) and our Colloquial Claims. Higher perplexity implies that sentences are far from the styles of dialogues.

FEVER		Colloquial Claims	
film	american	yes	know
born	released	actually	yeah
series	award	movie	film
stars	won	called	american
actor	united	born	like
directed	starred	heard	yea
television	album	released	played
world	worked	won	oh
movie	states	actor	award
john	played	people	world
written	appeared	series	album
role	character	great	directed

Table 5: Comparison between the top-20 tokens of FEVER and our Colloquial Claims.

Table 5 also compares the top-20 frequent tokens in the claims from FEVER and our dataset. The most frequent tokens in FEVER's claims are mostly fact-related words, such as "*american*", "*released*", and "*born*". On the other hand, the claims in our Colloquial Claims also have tokens that frequently appear in conversations, such as "*know*", "*actually*", "*like*", and "*oh*".

4.2 Qualitative Comparisons

We conduct human evaluation via Amazon Mechanical Turk to investigate the effectiveness of our filtering pipeline. We random sample 100 data instances from our Colloquial Claims and compare between survived and removed candidates. Each instance is rated by three unique human annotators.

To evaluate the overall quality of our generated claims, we ask human users to evaluate humanness in 4-point scale: "*Do you think this sentence is from a bot or a human?*". We compare them with responses from WoW and FEVER on humanness.

We also conduct NLI on the claims from our Colloquial Claims and FEVER to evaluate the label mappings. Users are instructed to classify claims into three veracity labels given the gold evidence: SUPPORTED, REFUTED, NOTENOUGHINFO.

	Humanness	Human NLI
Wizard of Wikipedia	3.12	-
FEVER	2.57	0.96
Removed Claims	2.95	0.50
Survived Claims	2.94	0.82

Table 6: Human evaluation results comparing the humanness and NLI performance between responses from Wizard of Wikipedia, claims from FEVER, our removed claims and survived claims.

Table 6 summarizes the averaged humanness and human NLI scores. Since the responses in WoW are from real dialogues, we can observe they have the highest humanness score. Interestingly, our generated claims are evaluated to be better than human-generated claims in FEVER, in terms of humanness. We suspect that this is due to the colloquialism of our generated claims.

The survived claims have more accurate label mappings with the evidence, compared to removed candidates. It is thanks to the bidirectional NLI filter that removes the candidate claims that are semantically different from the original claims. Table 7 shows some examples comparing our generated claims to the original FEVER claims.

FEVER: Google Search displays movie showtimes. **Colloquial Claim**: I can try google search to see what movie to watch and get show times!
FEVER: Unison (Celine Dion album) was originally released by Atlantic Records. **Colloquial Claim**: I remember the Celine Dion album titled Unison. It was released by Atlantic Records.
FEVER: Firefox is a desktop browser. **Colloquial Claim**: Yes, I use something called firefox for my desktop browser.
FEVER: Kung Fu Panda was released in theaters in 2006. **Colloquial Claim**: Have you watched Kung Fu Panda? It came out in 2006.
FEVER: San Francisco Bay Area contains many airports. **Colloquial Claim**: Sure, and yes there are lots of Bay Area airports!
FEVER: Brthday Song's (2 Chainz song) producer was Mike Dean. **Colloquial Claim**: Do you listen to Birthday Song by 2 Chainz ? It was produced by Mike Dean.

Table 7: Examples of generated colloquial claims for the original FEVER claims.

5 Experiments and Analysis

We conduct experiments on our curated colloquial claims to see how they impact existing fact checking systems.

5.1 Experimental Setting

Datasets. FEVER (Thorne et al., 2018) consists of three steps of fact checking pipeline: document retrieval, evidence selection, and claim verification. Based on selected evidence, the claims are classified into three classes of veracity: SUPPORTED, REFUTED, NOTENOUGHINFO. The Colloquial Claims is our generated dataset based on FEVER with claims in the colloquial style.

Metrics. FEVER fact checking uses two performance scores: label accuracy and FEVER-score. Label accuracy is the claim verification performance of the fact checking system. The FEVER-score is a more complicated evaluation regarding the whole pipeline. Following the FEVER challenge[2], a claim verification is evaluated as correct if the system retrieves at least one complete set of ground-truth evidence sentences and also classifies the claim correctly. For the evidence sentences, we evaluate the first 5 sentences retrieved from the system. We also report the recall for retrieved documents and selected evidence sentences.

5.2 Fact-Checking Baselines

We run experiments on six combinations of the fact-checking system according to the steps. For each dataset evaluation, we finetuned the system on the respective dataset.

Document Retrieval. We test three types of approaches: (1) oracle, (2) term-matching, and (3) similarity search with dense representation. First, the oracle always returns five evidence sentences including the gold evidence. Second, the WikiAPI[3], following Hanselowski et al. (2018), retrieves Wikipedia documents by matching words in the claim through a python library. Third, Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) retrieves documents via similarity search with BERT embeddings trained by metric learning.

[2] https://fever.ai/2018/task.html

[3] We adopt the implementation by Hanselowski et al. (2018) at https://github.com/UKPLab/fever-2018-team-athene

Dataset	Document Retrieval +Evidence Selection	Document Recall	Evidence Recall	Veracity Classification			
				KGAT(BERT)		KGAT(CorefBERT)	
				Label Accuracy	FEVER score	Label Accuracy	FEVER score
FEVER	Evidence Oracle	-	-	69.7	-	77.5	-
	WikiAPI + BERT	90.0	85.3	67.5	62.4	73.8	69.5
	Dense Passage Retrieval + BERT	84.0	81.8	62.9	55.4	61.1	52.4
Colloquial Claims	Evidence Oracle	-	-	57.3	-	67.7	-
	WikiAPI + BERT	72.2	73.4	53.2	43.6	60.9	52.4
	Dense Passage Retrieval + BERT	79.6	77.4	51.2	41.5	61.0	55.4

Table 8: Performance comparison of six fact checking system configurations with evidence oracle, WikiAPI, Dense Passage Retrieval (Karpukhin et al., 2020), BERT (Devlin et al., 2019), CorefBERT (Ye et al., 2020) with Kernel Graph Attention Network (KGAT) (Liu et al., 2020) on FEVER (Thorne et al., 2018) and our Colloquial Claims.

Evidence Selection. WikiAPI and DPR both use BERT (Devlin et al., 2019) to encode sentences and sort them out from the documents.

Claim Verification. We test two approaches: (1) BERT and (2) CorefBERT (Ye et al., 2020), which is one of the best performing methods on FEVER. The CorefBERT pretrains BERT to better capture the coreference information in text. We also apply kernel graph attention network (KGAT) (Liu et al., 2020) on BERT and CorefBERT for fine-grained attention using evidence graphs. More details can be found in Appendix.

5.3 Experimental Results

Table 8 compares the performance of fact checking systems on FEVER and our Colloquial Claims. Both label accuracy and FEVER-score significantly decrease for all systems on our Colloquial Claims, compared to FEVER. The WikiAPI+BERT+KGAT(CorefBERT) system performs on par with best performing models for FEVER by label accuracy of 73.8%. However, it degenerates on the colloquial dataset with the label accuracy of 60.9%. We remind that our Colloquial Claims shares the same document pool, annotated evidence sentences, and similar semantics with claims from FEVER. Thus, it is the difference in the claim's style that makes the fact checking systems fatally degenerate.

The WikiAPI, used in many fact checking systems (Hanselowski et al., 2018; Chernyavskiy and Ilvovsky, 2019; Stammbach and Neumann, 2019; Zhou et al., 2019; Liu et al., 2020), shows superior performance than DPR on the FEVER dataset, with document recall of 90.0%. On Colloquial Claims, however, it crashes down to 72.2%. Meanwhile, the DPR shows more robust document retrieval on Colloquial Claims than WikiAPI.

Apart from document retrieval and evidence selection, we can also observe performance decrease in the systems with evidence oracles. This indicates that claim verification is also more difficult on Colloquial Claims.

5.4 Challenges in Colloquial Claims

We analyze the causes of degeneration in document retrieval and claim verification in relation to the colloquial traits. We compare three document retrieval methods along with the oracle: WikiAPI, DrQA (Chen et al., 2017), and Dense Passage Retrieval (DPR). DrQA is another variation of term-matching method based on TF-IDF. Table 9 shows the titles of ten most documents by each retriever.

Filler Words Unnecessary of Fact Checking. In colloquial language, claims are not always composed of factual remarks requiring verification. Filler words (*e.g.* "*I see*", "*yeah, like*") are also frequently mixed in the utterances, as shown in Table 5. Hence, our Colloquial Claims requires systems to partition the parts that affect veracity from the ones that do not. However, Table 9 shows that word-matching retrieval systems, such as WikiAPI and DrQA, are vulnerable to those insignificant parts. They naively retrieve filler word related documents very frequently.

Minding the Context. Considering the context inside the sentence is essential for verifying colloquial claims. Lexical variation and polysemy is common in colloquial language. Such variations and ambiguity are tolerable because common context flows in the utterance. For example, in the colloquial claim of "*Niko Coster-Waldau is also the host of the show. He was with Fox at one point.*", it is easy to see the word "*Fox*" stands for "*Fox Broad-*

Oracle	WikiAPI
Pakistan	It
Pocahontas	I
SpongeBob	You
Far from the Madding Crowd	Yes (band)
Samsung	Yes (album)
Two and a Half Men	He
Elizabeth of York	That
Ice-T	They
Spiderman	There There (novel)
Sausage Party	HES
DrQA	**DPR**
Heroes of Russia	Minor League
Yeah Yeah	Beverly Hillibillies
Yea	Ed and Lorraine Warren
H*** Yeah	Benjamin Franklin
Yea (football club)	Yin and Yang
Stefanie Drootin	Hunger Games (film)
Minor League	Sausage Party
Video Games	Ice-T
Google Search	Mormons
Google Apps	Burj Khalifa

Table 9: Comparison of the titles of the top-10 retrieved documents between oracle, WikiAPI, DrQA and DPR.

casting Company" based on the context. However, it is well known that simple term-matching methods cannot capture such context (Karpukhin et al., 2020). Thus, we observe that systems instead simply retrieve the document of "*fox*". Also, Table 9 shows another example of contextless retrieval. The document "*Yes (band)*", "*There There (novel)*", and "*Yea (football club)*" are naively retrieved by the systems, due to simple filler words in colloquial claims.

Overcoming the Colloquial Traits. Methods based on TF-IDF or word-matching are good at recognizing core keywords, but suffer at capturing the rich semantics of context. On the other hand, the DPR, a similarity search method based on dense embeddings, shows promising results. Results in Table 9 illustrate that DPR is able to ignore the context-irrelevant entities and focus more on fact-related entities. Compared to other retrieval methods, the ten most retrieved documents from DPR does not contain any filler words. Since filler words are irrelevant to the veracity of colloquial claims, the DPR learns their insignificance. Therefore, dense representation can be important for making fact-checking systems to be robust on claims in dialogues.

6 Related Work

Fact Checking and Verification. The need for claim verification has led to annotated fact checking datasets (Thorne et al., 2018; Baly et al., 2018; Augenstein et al., 2019; Jiang et al., 2020; Wadden et al., 2020; Chen et al., 2020). Recent works deploy adversarial attacks against fact checking systems (Thorne et al., 2019a,b; Niewinski et al., 2019; Atanasova et al., 2020b) and attempt to improve the system through generation (Atanasova et al., 2020a; Goyal and Durrett, 2020; Fan et al., 2020). Existing works tend to focus on verifying news or Wikipedia. However, verifying facts is not limited to such formal texts. Compared to previous works, we focus on verifying claims in the dialogue domain, which resembles more daily life situations.

A special case of fact verification is rumour detection. Its goal is to determine the veracity of rumours from social media (Li et al., 2019). The rumour is classified based on the reactions of chained messages (Gorrell et al., 2019). The procedure and characteristics of rumour detection is quite different from the fact checking pipeline (Gorrell et al., 2019). In our task, we verify the claims based on factuality from the related documents, rather than stances of the comments.

Safety in Open-domain Dialogue. Recently, much work has studied safety issues of machine dialogue agents in several aspects. Wulczyn et al. (2017) attempt to detect personal attacks in Wikipedia talk pages. Henderson et al. (2018) note the axes of bias, adversarial examples, privacy and safety, and propose that the community should aim to provide conditional safety guarantees. Khatri et al. (2018) train a sensitive language detector to evaluate the utterances in a chatbot dataset. Dinan et al. (2019a) propose a framework for dialogue agents to be robust to malicious human attacks. Other works have attempted to mitigate biases, such as gender bias (Dinan et al., 2020a) and racial bias (Sap et al., 2019). Recently, Tran et al. (2020) modify BERT (Devlin et al., 2019) to detect hatespeech. Xu et al. (2020) introduce a method to distill safety standards into the generative dialogue agent.

Previous works cover a wide range of dialogue safety, yet the risk of disinformation and misinformation remain understudied. In this work, we extend dialogue safety to cover verification of responses with false information.

7 Conclusion

This work aimed to open up new discussions in the intersection of fact checking and dialogue safety. In order to study how existing fact checking systems behave on claims in dialogues, we curate colloquial claims by transferring the styles of claims in FEVER (Thorne et al., 2018) to colloquialism. We leverage BART (Lewis et al., 2020) and Wizard of Wikipedia (WoW) (Dinan et al., 2019b). We finetune BART to generate the wizard's responses with knowledge sentences from WoW. Then, we input FEVER claims to generate claim-grounded utterances. We oversample candidate claims and apply filters to compensate quality. We showed that existing fact checking systems well-performing on FEVER degenerate on colloquial claims. We found that the document retriever is the weakest spot in the system which is even vulnerable to filler words. We compared the characteristic differences between claims in formal style and ones in colloquialism . An important future direction will be building a dialogue dataset for fact checking.

Acknowledgements

We would like to thank Jinseo Jeong and Myeongjang Pyeon for their valuable comments. We also thank the anonymous reviewers for their thoughtful suggestions on this work. This research was supported by Samsung Research Funding Center of Samsung Electronics under project number SRFC-IT2101-01. Gunhee Kim is the corresponding author.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating Fact Checking Explanations. In *ACL*.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating Label Cohesive and Well-Formed Adversarial Claims. In *EMNLP*.

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-IJCNLP*.

R. Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez i Villodre, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *NAACL-HLT*.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and Consumer Safety Risks when using Conversational Assistants for Medical Information: an Observational Study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research*, 20(9):e11510.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial Filters of Dataset Biases. In *ICML*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *ACL*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *ICLR*.

Anton Chernyavskiy and Dmitry Ilvovsky. 2019. Extract and Aggregate: A Novel Domain-Independent Approach to Factual Data Verification. In *FEVER*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *EMNLP*.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack. In *EMNLP-IJCNLP*.

Emily Dinan, Verena Rieser, Alborz Geramifard, Zhou Yu, and Dilek Hakkani-Tür. 2020b. Safety for Conversational AI Workshop. https://safetyforconvai.splashthat.com/.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *ICLR*.

Mihail Eric and Christopher D Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *SIGDIAL*.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating Fact Checking Briefs. In *EMNLP*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A Knowledge-Grounded Neural Conversation Model. In *AAAI*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Interspeech*.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2019. RumourEval 2019: Determining Rumour Veracity and Support for Rumours. In *NAACL-HLT SemEval workshop*.

Tanya Goyal and Greg Durrett. 2020. Evaluating Factuality in Generation with Dependency-level Entailment. In *Findings of EMNLP*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-Athene: Multi-sentence Textual Entailment for Claim Verification. In *FEVER*.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In *AIES*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *ICLR*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In *EMNLP*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*.

Chandra Khatri, Behnam Hedayatnia, Rahul Goel, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Detecting offensive content in open-domain conversations using two stage semi-supervision. In *NeurIPS ConvAI Workshop*.

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*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *ACL*.

Quanzhi Li, Q. Zhang, L. Si, and Yingchi Liu. 2019. Rumor Detection on Social Media: Datasets, Methods and Opportunities. In *EMNLP NLP4IF Workshop*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In *ACL*, pages 7342–7351.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A Dialog Research Software Platform. *arXiv:1705.06476*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *EMNLP*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *AAAI*.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative Enhanced Model for Adversarial Attacks. In *EMNLP FEVER Workshop*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *ACL System Demonstration*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by Reading: Contentful Neural Conversation with On-Demand Machine Reading. In *ACL*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for Building an Open-Domain Chatbot. *arXiv:2004.13637*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *AAAI*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *ACL*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the fever shared task. In *FEVER*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019a. Evaluating Adversarial Attacks Against Multiple Fact Verification Systems. In *EMNLP*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019b. The FEVER2.0 Shared Task. In *EMNLP FEVER Workshop*.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Serim Park. 2020. HABERTOR: An Efficient and Effective Deep Hatespeech Detector. In *EMNLP*.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *EMNLP*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0. *LDC*, 23.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A Broad-coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL-HLT*.

M. Wolf, K. Miller, and F. Grodzinsky. 2017. Why We should have Seen that Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. *SIGCAS Computers and Society*, 47:54–64.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-Art Natural Language Processing. *arXiv:1910.03771*.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness. In *ACL*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *TheWebConf*.

Jing Xu, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for Safety in Open-domain Chatbots. *arXiv:2010.07079*.

Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai, Liran Wang, Wei Wang, and Zhoujun Li. 2020. StyleDGPT: Stylized Response Generation with Pre-trained Language Models. In *Findings of EMNLP*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *EMNLP*.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2018. Augmenting End-to-end Dialog Systems with Commonsense Knowledge. In *AAAI*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-Scale Generative Pre-Training for Conversational Response Generation. *arXiv:1911.00536*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *ACL*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *ACL*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A Dataset for Document Grounded Conversations. In *EMNLP*.

A Implementation Details of AFLITE

We use adversarial filtering method AFLITE (Sakaguchi et al., 2020; Bras et al., 2020) to select top-k candidate claims which are most difficult to discriminate from responses in Wizard of Wikipedia (WoW) (Dinan et al., 2019b). The algorithm takes as input the original WoW and Colloquial Claims, then returns each filtered dataset. AFLITE comprised with two steps: (i) precomputing phase and (ii) filtering phase.

In precomputing phase, we randomly sample 10% of instances from WoW and Colloquial Claims to fine-tune RoBERTa-large. We then use fine-tuned RoBERTa to pre-compute embeddings for the rest of the instances as the input for the filtering phase. We discard samples used for fine-tuning from the final dataset.

In filtering phase, we use an ensemble of linear classifiers to iteratively discard easily distinguishable instances. At each iteration, we train 32 linear classifiers on different random partitions of the data and collect their predictions on their rest of the instances. For each instance, we compute its score as the ratio of correct predictions over the total number of predictions, and remove top-n instances whose score is above threshold 0.75. We remove top-1000 instances among the entire WoW, and top-2 instances for each candidate sets in Colloquial Claims. We repeat this process until we have less than 3 instances for each candidate set or scores in candidate set are below the threshold.

B Other Implementation Details

For finetuning BART-large (Lewis et al., 2020) on Wizard of Wikipedia (Dinan et al., 2019b) dataset, we use the ParlAI framework[4] (Miller et al., 2017) with default hyperparameters. We use RoBERTa-large (Liu et al., 2019) from HuggingFace's Transformers[5] (Wolf et al., 2019) to implement bidirectional NLI, and named entity recognition module from Stanza[6] (Qi et al., 2020) to extract named entities from generated claims and claims in FEVER. We use official code from the authors to implement KGAT and BERT evidence selector[7] (Liu et al., 2020), CorefBERT[8] (Ye et al., 2020), DPR[9] (Karpukhin et al., 2020), and WikiAPI document retriever[10] (Hanselowski et al., 2018). We finetune CorefBERT-base for CorefBERT and BERT-base for BERT evidence selector, BERT claim verifier and DPR. We use default hyperparameters for all the experiments.

For DPR, we use preprocessed English Wikipedia dump from FEVER 1.0[11] as the source documents for retrieval, which contains 25,248,398 evidence sentences from 5,396,106 documents. We use documents from top-10 retrieved evidences as a document retrieval result, which contains 7.2 documents in average.

All the experiments are run on up to 8 NVIDIA Quadro RTX 6000 GPUs.

C Claim Examples

FEVER (REFUTED):
Dave Gibbons has always been unable to write.
Colloquial Claim:
For some reason Dave Gibbons has always been unable to write.

FEVER (SUPPORTED):
Phillip Glass has written eleven concertos.
Colloquial Claim:
I'd like to suggest Phillip Glass. He has written a total of eleven concertos!

FEVER (REFUTED):
Planet Hollywood Las Vegas is owned by Leonardo DiCaprio.
Colloquial Claim:
Oh okay well if you ever come to LV go to the Planet Hollywood building, its owned by Leonardo DiCaprio.

FEVER (NOTENOUGHINFO):
General Motors had only one automotive-component.
Colloquial Claim:
That company used to be called General Motors, General Motors had only one automotive-component.

FEVER (REFUTED):
Steve Ditko studied art at the Cartoonist and Illustrators School.

[4] https://parl.ai/
[5] https://huggingface.co/transformers/
[6] https://github.com/stanfordnlp/stanza
[7] https://github.com/thunlp/KernelGAT
[8] https://github.com/thunlp/CorefBERT
[9] https://github.com/facebookresearch/DPR
[10] https://github.com/UKPLab/fever-2018-team-athene
[11] https://fever.ai/resources.html

Colloquial Claim:
That's cool. I read that Steve Ditko studied at the Cartoonist and Illustrators School.

FEVER (NOTENOUGHINFO):
Arjit Singh goes unmentioned in the Indian media.
Colloquial Claim:
I heard Arjit Singh doesn't get much attention in the Indian media.

FEVER (SUPPORTED):
The Cry of the Owl is based on Patricia Highsmith's eighth novel "Push".
Colloquial Claim:
Yep! In fact, the movie Cry of the Owl is based on a Patricia Highsmith book called Push!

FEVER (SUPPORTED):
Justin Chatwin is an actor.
Colloquial Claim:
In case you didn't already know Justin Chatwin is an actor.

FEVER (REFUTED):
Dreamer (2005 film) was directed by Michael Bay only.
Colloquial Claim:
It is true! There was even a Michael Bay film called Dreamer released in 2005.

FEVER (NOTENOUGHINFO):
Harvard University is a commuter school.
Colloquial Claim:
I hear that Harvard is a commuter school.

FEVER (REFUTED):
In 2015, among Americans, 44% of adults had consumed alcoholic drink in the last month.
Colloquial Claim:
Yes, in 2015, a shocking 44% of adults reported having consumed alcohol in the last month.

FEVER (SUPPORTED):
Sands Hotel and Casino started in 1952 as a casino with 200 rooms.
Colloquial Claim:
You will have to go to the Sands Hotel and Casino to gamble! It was founded in 1952 with 200 rooms.

FEVER (SUPPORTED):
Zoe Saldana's birth year was 1978.
Colloquial Claim:
Are you familiar with Zoe Saldana? Her birth year was 1978!

FEVER (NOTENOUGHINFO):
Iraq is in the Group of 15.
Colloquial Claim:
I know that Iraq is in the group of 15.

FEVER (REFUTED):
Bala has no experience directing.
Colloquial Claim:
Not really a director. And Bala does not have any experience in directing at all!

FEVER (SUPPORTED):
Padua is the political hub of the area.
Colloquial Claim:
Well, I know that Padua is considered the political hub of the area.

FEVER (SUPPORTED):
Sensitive Skin's first series aired on ABC TV.
Colloquial Claim:
I know that the first episode of Sensitive Skin aired on ABC!

FEVER (NOTENOUGHINFO):
Baadshah was dubbed into Portuguese.
Colloquial Claim:
yeah Baadshah was dubbed in portuguese as well.

FEVER (SUPPORTED):
The Times has been printed since 1785.
Colloquial Claim:
Well since 1785 the times has been around!

FEVER (REFUTED):
The IPhone 4 was designed by cats.
Colloquial Claim:
The Iphone 4 actually was designed by cats. Can you believe that? It was designed by cats.

FEVER (SUPPORTED):
Little Dorrit is a novel by Charles Dickens written in the 1850s.
Colloquial Claim:
Yeah. The little dorrit was written by Dickens way back in the 1850's.

FEVER (SUPPORTED):

Anne Boleyn is an influential person that was mentioned in many artistic and cultural work.
Colloquial Claim:
Yes I think so. Anne Boleyn was really influential in many different arts and cultural works.

FEVER (NOTENOUGHINFO):
Bank of America provides products and blankets.
Colloquial Claim:
I understand that. One company that provides a lot of blankets is Bank of America.

FEVER (NOTENOUGHINFO):
Amancio Ortega was born on a boat.
Colloquial Claim:
His real name is Amancio Ortega and he was born on a boat. Interesting fact!

FEVER (REFUTED):
Annie was released in 2016.
Colloquial Claim:
I heard that the movie Annie was released in 2016.

FEVER (SUPPORTED):
2 Hearts is a song by Minogue.
Colloquial Claim:
Yes the song 2 hearts was by kylie minogue.

FEVER (NOTENOUGHINFO):
Ice-T made a hip-hop album in 1999.
Colloquial Claim:
No, but Ice-T made a hip-hop album in 1999.

FEVER (REFUTED):
Barbarella was directed solely by George Lucas.
Colloquial Claim:
In case you're curious, Barbarella was directed by George Lucas.

FEVER (SUPPORTED):
Jon Hamm received Primetime Emmy Award nominations for his performances in Mad Men.
Colloquial Claim:
You should! Especially Jon Hamm's performance in Mad Men! It earned him Primetime Emmy nominations!

FEVER (SUPPORTED):
Alvin and the Chipmunks's director was Tim Hill.
Colloquial Claim:
Yes they did. I'm reminded of Alvin and the Chipmunks. Tim Hill directed the animated film.

FEVER (NOTENOUGHINFO):
Daenerys Targaryen is the last surviving member of House Targaryen.
Colloquial Claim:
Yep! Daenerys Targaryen is the only remaining member of the Targaryen family!

FEVER (NOTENOUGHINFO):
In North America, Warcraft was released by Universal Pictures.
Colloquial Claim:
Well Warcraft was released by Universal Pictures.