DialFact: A Benchmark for Fact-Checking in Dialogue

Anonymous ACL submission

Abstract

Fact-checking is an essential tool to mitigate the spread of misinformation and disinformation. We introduce the task of fact-checking in dialogue, which is a relatively unexplored area. We construct DialFact, a testing benchmark dataset of 22,123 annotated conversational claims, paired with pieces of evidence from Wikipedia. There are three sub-tasks in DialFact: 1) Verifiable claim detection task distinguishes whether a response carries verifiable factual information; 2) Evidence retrieval task retrieves the most relevant Wikipedia snippets as evidence; 3) Claim verification task predicts a dialogue response to be supported, refuted, or not enough information. We found that existing fact-checking models trained on non-dialogue data like FEVER (Thorne et al., 2018) fail to perform well on our task, and thus, we propose a simple yet data-efficient solution to effectively improve fact-checking performance in dialogue. We point out unique challenges in DialFact such as handling the colloquialisms, coreferences and retrieval ambiguities in the error analysis to shed light on future research in this direction[1].

1 Introduction

Misinformation online can have deleterious consequences to our society, especially during public health crises like the COVID-19 pandemic. False and outdated information can be spread not only by humans but also by automatic agents as generative models have shown remarkable progress recently (Adiwardana et al., 2020; Xu et al., 2021). These systems are not perfect, they can either generate hallucinated and imperfect information, or they can be abused to automatically generate false claims and spread misinformation at a massive scale. Fact verification tools are thus necessary in the current information age to tackle the spread of misinformation propagated.

[1]Data and code are available at `github.com/xyz`

Dialogue Context: I have family in Ireland! Have you ever been there?
Evidence: Ireland is an island in the North Atlantic.

Non-Verifiable Response: I haven't been but want to!
Verifiable Supported Response: I haven't. It is an island in the north Atlantic right?
Verifiable Refuted Response: I haven't been. Isn't it somewhere in north Pacific?
Verifiable NEI Response: I haven't been. I heard it's the most popular tourist location in Europe!

Figure 1: Dialogue fact-checking involves predicting if a response should be considered a Verifiable claim, followed by finding relevant evidence, and finally predicting if the it is Supported, Refuted or NEI.

Fact-checking was introduced in Wang (2017); Thorne et al. (2018) and since then a growing body of research has explored and suggested various tasks and resources to address the challenges in this area. Fact-checking has been explored in medium such as Wikipedia passages, tables, social media and news articles (Guo et al., 2021). However, there is no data available for fact-checking in dialogue, and related work mainly focuses on improving factual consistency in knowledge-grounded response generation (Honovich et al., 2021; Rashkin et al., 2021; Shuster et al., 2021).

Verifying factual correctness of claims in dialogue poses new challenges to both dataset construction and modeling. Claims in existing datasets are from formal sources such as news articles and they are generally succinct and formal. In contrast, claims in dialogue are often informal and sparse in factual content. Furthermore, dialogue utterances often include personal opinions, slang, and colloquialisms which need to be distinguished from factual information. Another challenge in dialogue fact-checking is that ellipsis and coreference occur frequently which make utterances incomplete and ambiguous (DeVault and Stone, 2007). Although humans can easily understand utterances with references or absent information based on the dialogue context and their reasoning skills, a fact-checking system may need to model this behavior explicitly.

We introduce the task of fact-checking in dialogue and propose an evaluation dataset DIALFACT. It has 22,123 annotated conversational claims, 10,355 in the validation set, and 11,768 in the test set. An example is shown in Figure 1. DIALFACT has three sub-tasks: 1) Verifiable claim detection aims to distinguish responses that do not contain verifiable factual information, such as "I haven't been but want to!" in Figure 1. 2) Evidence retrieval involves selecting the most relevant knowledge snippets from Wikipedia which can verify the response. 3) Claim verification aims to classify if a response is supported, refuted, or does not have enough information to verify the response given the dialogue history and the retrieved evidence.

DIALFACT consists of both human-written and machine-generated claims based on the Wizard of Wikipedia (Dinan et al., 2019) dialogue dataset. Each response claim and its evidence sentences from Wikipedia are annotated by crowd workers and we perform rigorous quality checks on the annotations. For fact verification, we propose creation of weakly-supervised training data by leveraging techniques such as negation, entity swapping, language model mask-and-fill, and knowledge-grounded generation. We establish baseline model performance on this task, and point out the weaknesses of fact-checking models. Our analysis show that this is a non-trivial task with challenges remaining for future work.

2 Related Work

Fact Verification The spread of false information online has led to a growing body of research exploring automatic fact-checking. Thorne et al. (2018) and subsequent works (Wenhu Chen et al., 2020; Jiang et al., 2020; Nørregaard and Derczynski, 2021; Aly et al., 2021) introduced fact extraction and verification datasets verifiable against pieces of evidence from Wikipedia articles. Fact-checking has been explored in variety of medium such as Wikipedia based claims (Schuster et al., 2021), claims over tables (Aly et al., 2021), scientific claims (Wadden et al., 2020), and social media claims (Nakov et al., 2021). However, fact-checking in dialogue is still an unexplored area. Kim et al. (2021) explored fact-checking for colloquial claims, curated by converting FEVER claims into colloquial style. Although closely related to our work, colloquial claims is not a dialogue dataset, only contains verifiable claims, and does not have dialogue contexts for claims. In DIALFACT, on the other hand, both evidence retrieval and claim verification are more challenging as they require resolving ambiguities and coreferences from the dialogue context.

Consistency in Dialogue Neural dialogue systems grounded on knowledge sources such as Wikipedia (Dinan et al., 2019), knowledge graphs (Wu et al., 2019) or snippets from the internet (Komeili et al., 2021) have garnered interest in recent years. Despite generating plausible and engaging responses, existing models still hallucinate invalid information (Roller et al., 2021). Ensuring safety and consistency in dialogue response generation is thus an actively explored area (Rashkin et al., 2021; Shuster et al., 2021). Some recent works have proposed evaluation metrics and benchmarks for factual consistency in knowledge grounded response generation (Honovich et al., 2021; Dziri et al., 2021). Our work instead focuses on fact-checking in dialogue for both human and machine-generated responses, and involves additional tasks of verifiable claim detection and evidence retrieval.

Synthetic datasets Synthetic dataset construction has been shown to improve robustness of evaluation models (Gupta et al., 2021; Ghazarian et al., 2021) and improve the complexity of test sets (Sakaguchi et al., 2021; Feng et al., 2021). Synthetic claims have been explored in fact-checking to create hard test sets. Several participants in the FEVER 2.0 breakers phase (Niewinski et al., 2019; Hidey et al., 2020; Atanasova et al., 2020) proposed approaches for automatically generated adversarial claims. Recently, Jiang et al. (2020) created complex multi-hop claims using word substitutions, Saakyan et al. (2021) used Bert based token-infilling to created refuted claims, and Schuster et al. (2021) created synthetic revisions to Wikipedia sentences to improve fact-checking robustness. Our work also introduces techniques to create synthetic claims in the context of dialogue fact-checking.

3 Task Background

Let a conversation context consist of a list of utterances $C = \{u_1, u_2, ..., u_n\}$. The task is to perform fact-checking on the last utterance of the conversation u_n, henceforth called claim c. Fact-checking claims in conversations is a pipeline that consists of several steps. First, the system needs to decide whether a response is VERIFIABLE or

NON-VERIFIABLE. We define them as follows: **NON-VERIFIABLE:** The claim contains no verifiable factual information. It includes claims with personal opinions or personal information. **VERIFIABLE:** The claim contains at least one factual information verifiable against a background corpus (Wikipedia in this task).

Next, the system should retrieve documents from the background corpus and select relevant evidence sentences from the documents. Finally, the system should predict whether the claim belongs to one of the following three categories: **SUPPORTED:** The response contains factual information which is valid in light of the evidence. **REFUTED:** The response contains factual information which is invalid in light of the evidence. **NOTENOUGHINFORMATION (NEI):** The response contains factual information which can not be validated (supported or refuted) with the evidence.

VERIFIABLE claims can be SUPPORTED, REFUTED, or NEI, and NON-VERIFIABLE claims are always NEI. We leverage the *Wizard of Wikipedia* (WoW) dataset (Dinan et al., 2019) as the base to build this task. WoW is a knowledge-grounded open-domain dialogue dataset with conversations between two speakers - a wizard who has access to background Wikipedia documents to deliver knowledge carrying responses, and an apprentice who plays the role of a curious learner. For each turn u_i, the wizard is shown a set of articles K_i retrieved from Wikipedia. The wizard either chooses a relevant knowledge sentence k_i from the set K_i, or chooses a *no sentence used* option to construct a response. For our fact-checking task, we additionally need claims which belong to REFUTED and NEI categories. We next describe the methodologies used to create claims from the valid and test splits of the WoW dataset.

4 Dataset Construction and Annotation

We use two approaches to create claim responses for DIALFACT: 1) Automatically generated claims, and 2) Human written claims to emulates claims created by dialogue systems and humans respectively. All claims are further annotated by crowd workers on Amazon Mechanical Turk (Mturk).

4.1 Automatically Generated Claims

In this approach, we use automatic methods to create claims for all categories either from scratch or by mutating the responses in WoW dataset.

4.1.1 Methods for claim generation

Negation We use the 42 rule-based transformations from Thorne et al. (2019) which apply to verb phrases of the claims to convert them to their negated versions by adding words like "not" or "no". It typically creates REFUTED claims.

Substitution We perform three types of substitutions: For 1) Context and knowledge-based entity substitution, we first run SpaCy NER tagging (Honnibal and Montani, 2017) on a response u_i from WoW. We then swap an entity in the response u_i with an entity from either its conversation context C or its background knowledge articles set K_i. An entity is only swapped if it is present in k_i, the original knowledge sentence to avoid swaps which do not change the facts. Entities are swapped within their types. For 2) Sense-based substitution, we swap an entity in u_i with an entity with a similar "sense" returned from the sense2vec (Trask et al., 2015) library. For 3) Adjective substitution, we substitute adjectives in a claim (ignoring adjectives related to emotions, such as "happy") with their WordNet (Miller, 1998) antonyms (for example *best* is replaced with *worst*). These operations typically create REFUTED claims.

Mask-and-Fill This method generates claims in two stages: 1) Mask salient words from the original claims, and 2) Substitute those words with their alternates using a language model. For masking salient words in the original response claims, we follow the procedure from Thorne and Vlachos (2021) and use the Neutrality Masker model from Shah et al. (2020). It predicts the tokens which upon masking are likely to cause a label flip from SUPPORTED to NEI. For step 2) we first train a T5-base model (Raffel et al., 2020) on the WoW dataset on the task of infilling masked tokens conditioned on evidence sentences. For training, the input sequence consists of concatenated evidence sentence k_i, dialogue context C, and the gold response with masked spans at random positions, and the output is the gold response. The model is thus trained to infill a masked response based on the provided evidence and the dialogue context. For generating response claims which belong to REFUTED or NEI categories, we use the following types of evidence sentences to condition the infilling: a) empty evidence, b) evidence sentences selected randomly from the knowledge article set K_i belonging to the original response, and c) evidence sentences from a Wikipedia article of an entity retrieved using sense2vec based on its similarity with the entities in the original response. Conditioning on such evidence lead to generation of claims which have factual details inconsistent with the original evidence.

Generation We fine-tune one of the best chit-chat dialogue systems, Blenderbot model (Roller et al., 2021), on the WoW dataset. The model takes the concatenation of the knowledge sentence k_i and the dialogue context C as input and it is trained to predict the tokens of the gold response. To generate new response claims, we condition the model on the three types of evidence described in the Mask-and-Fill approach. We use a high temperature (1.5) and nucleus sampling (Holtzman et al., 2020) with $p = 0.9$ during decoding to encourage the model to generate unexpected and non-contextual entities in the responses.

Final claim set creation Our target is to create a challenging and diverse test set for dialogue fact-checking. Using the aforementioned methods of claim generation, we get a set $R_c = \{r_1, r_2, ..., r_k\}$ of response claims for a dialogue context C. To select a final set of claims, we first remove any responses which do not have at least 3 words different from other responses in R_c, then filter out less fluent claims whose GPT-2 (Radford et al., 2019) perplexity scores are higher than 1.1 times the average perplexity scores of the responses in R_c. We then score the response claims using existing state-of-the-art models related to our task: namely Dialogue NLI (Welleck et al., 2019), Dialogue contradiction detection (Nie et al., 2021), FEVER based fact verification (Schuster et al., 2021) and fact-checking on colloquial claims (Kim et al., 2021). For each model, we calculate the entropy of the scores predicted for each label and rank the claims in R_c based on the sum of the entropy of the scores of all the models, which gives an estimate of the confusion or difficulty in classifying the claims. The top 4 responses from the ranked list are chosen as the final set of response claims for that context.

4.1.2 Evidence set creation

For each claim, a set of evidence sentences is first automatically created and then labelled by crowd workers. We first extract a set of named entities and noun phrases n_k from the following sources: the claim c, the dialogue context C, the original response u_i for the dialogue context in WoW, and the title of the knowledge articles K_i shown to the wizard for u_i. We use the MediaWiki API[2] to find a set of relevant Wikipedia pages P_c for n_k. We then create a set of candidate sentences with the first 10 sentences of each page in P_c. Finally, we use two methods - SpaCy's word2vec similarity[3] and BM25 similarity[4] to rank the top 10 evidence sentences using each method. We then combine the non-overlapping evidence from both methods to create the final evidence set e_c for each claim c. We add the knowledge sentence k_i associated with the original response in the WoW dataset if it is not already present in e_c.

4.1.3 Claim and Evidence Annotation

We carry out the annotations of the claims and evidence on the Mturk platform in 3 rounds. The screenshot of the annotation UI is shown in Figure 3 of the Appendix. In each round a worker sees the claim c, its dialogue context C, and its associated evidence sentences e_c. Workers have to perform 3 tasks: First, they select if the claim is VERIFIABLE or NON-VERIFIABLE. Second, they select one or more evidence sentences related to the response claim. In case the set of evidence shown is not enough to decide the label of the response, or if they choose NEI, they are instructed to search Wikipedia and add relevant additional evidence sentences in the interface. For NEI claims they are instructed to add evidence sentences which are most related to the claim. Third, they choose the category of the response - SUPPORTED, REFUTED, or NEI. For NON-VERIFIABLE claims, NEI is auto-selected. Since automatically created responses can have grammatical or coherence related issues, in the first round of labeling, annotators are asked to edit a response to make it appropriate to the context if needed, or mark a response as incoherent, in which case it is removed from further rounds (We dropped 5% of incoherent claims). In the second and third rounds we gather 2 additional annotations for each claim. We select the label which has the majority vote among the set of 3 annotations across all rounds. The evidence set for each claim is the union of evidence annotated in any of the rounds.

4.2 Human Written Claims

Our dataset also consists of human written claims to cover lexical and stylistic patterns present in human-human conversations. The annotation is

[2] www.mediawiki.org/wiki/API:Main_page
[3] www.spacy.io/
[4] www.github.com/dorianbrown/rank_bm25

Validation					
	Supported	Refuted	NEI-Factual	NEI-Personal	Total
Generated	1841	1177	273	1256	4547
Written	1656	2316	1836	0	5808
Total	3497	3493	2109	1256	10355
Test					
	Supported	Refuted	NEI-Factual	NEI-Personal	Total
Generated	2664	1385	1105	1113	6183
Written	1493	2740	1268	0	5585
Total	4157	4125	2373	1113	11768

Table 1: Dataset statistics of DIALFACT for all categories and splits. *Generated* denotes automatically generated and *Written* denotes human written claims.

carried out in 3 rounds. *In the first round*, we instruct crowd workers to write VERIFIABLE factual responses conditioned on dialogue context and a set of evidence sentences for a pre-specified label l_c - one of SUPPORTED, REFUTED, or NEI. Workers were provided detailed examples and instructions for the task such as "Avoid using negation words such as do not, no for Refuted claims" (Appendix D). The evidence set for each claim is constructed using the method described in section 4.1.2. *In the second round*, we use the claim labeling interface from section 4.1.3 to gather labels for the claims collected in the first round. For any claim which is not labeled in the second round with the original label l_c, we gather a third round of annotations. If the label in the third round does not match l_c, we drop that claim from the dataset. We drop about 7% of the human written claims.

4.3 Dataset Statistics

We present the dataset statistics in Table 1. The dataset consists of total 22,123 claims across validation and test set with balanced SUPPORTED and REFUTED claims. Test set contains claims for 3,760 dialogue contexts with an average of 3.1 claims per context, and validation contains claims for 3,738 contexts with an average of 2.8 claims per context. The average number of tokens per claim is 22.0 in test set and 20.0 in validation set. Average number of evidence per claim is 1.3 in the test set and 1.1 in the validation set. We show some sample instances in Table 13 in the Appendix.

4.4 Quality Control

Annotators: We hire workers on Mturk with with at least 5000 HITS done and an acceptance rate of 95% or above. Workers have to first pass a qualification test where they are shown the task instructions, label definitions, and multiple examples and the explanations for each label. Then they are asked to label or write 12 claims. Using these qualification tests, we get a final set of 87 workers for the main data collection stage (Appendix D).

Quality checks Annotations were carried out in batches over multiple weeks. We examined random samples to provide feedback to workers. Workers with poor annotations were either asked to retake a new qualification test or removed from further batches. We recollected annotations for data annotated by removed workers. We provide tooltips and examples during annotation, and we also added automatic checks to alert workers about issues such as too short responses, no evidence selected, and copy-pasting evidence sentences as claims.

Data validation To evaluate inter-annotator agreement, we collected 2 extra rounds of annotations for 1200 claims for both automatically generated and human written claims, which is 10% of the data. Krippendorff's alpha value for category labels was 0.68 for human written claims and 0.58 for automatically generated claims, denoting moderate agreement. Krippendorff's alpha for VERIFIABLE versus NON-VERIFIABLE was 0.49, with a low-to-moderate agreement. The lower agreement is due to some claims like "Guns N' Roses was the greatest rock band of all time.", where it is difficult to judge if this is a personal opinion or a verifiable fact. In such conflicts, workers would still typically correctly label such ambiguous claims as NEI. We present lexical bias analysis in Appendix A.

5 Experiments

We propose new baselines and compare with existing models for three sub-tasks in dialogue fact-checking - 1) Verifiable claim detection, 2) Evidence retrieval, and 3) Claim verification.

5.1 Verifiable Claim Detection

We propose three simple baselines for verifiable claim detection. 1) *Lexical overlap* calculates the maximum word overlap between a claim and all evidence sentences after removing punctuation and stopwords using SpaCy. 2) *DNLI* uses the probability of the neutral class from the Dialogue Natural Language Inference model (Welleck et al., 2019). 3) *Lexical+DNLI* uses the sum of scores of both baselines and *Random* predicts each class with 50% probability. For all baselines, we mark a response as VERIFIABLE or NON-VERIFIABLE based on a

Baseline	Accuracy	Verifiable F1	Non-Verifiable F1
Random	50.0	64.2	19.2
Lexical	79.4	88.1	33.8
DNLI	82.1	89.9	37.1
Lexical+DNLI	**82.8**	**90.2**	**39.1**

Table 2: Accuracy and Macro F1 scores for Verifiable claim detection on the test set.

threshold value selected using validation data. We present the accuracy and individual F1 scores for both classes in Table 2. *Lexical+DNLI* performs the best and all baselines have low F1 scores for NON-VERIFIABLE claims.

5.2 Evidence Retrieval

Evidence retrieval consists of two steps: 1) Document Retrieval, 2) Evidence Sentence selection.

5.2.1 Document Retrieval

We test two methods for document retrieval: The first one is *WikiAPI*[5], which retrieves Wikipedia pages and is used in past fact-checking work (Hanselowski et al., 2018; Stammbach and Neumann, 2019; Liu et al., 2020). It uses the AllenNLP constituency parser (Gardner et al., 2018) to extract potential entities from the claims. Then it feeds the entities as queries through the MediaWiki API[2] and returns up to three Wikipedia pages per query. For each Wikipedia page, we query the KILT (Petroni et al., 2021) knowledge source to get the first 5 paragraphs of the page. We create two versions of this method: a) *Wiki-ctx* which concatenates the last two turns of the dialogue context with the response claim before document retrieval and b) *Wiki-claimonly* - which uses just the claim. The second method is *Dense Passage Retrieval (DPR)* (Karpukhin et al., 2020), a dual encoder based model which retrieves documents using BERT (Devlin et al., 2019) trained by metric learning. We create three versions of this method: a) *DPR-original*, which uses the original DPR trained on question-answering tasks, b) *DPR-WoWft-claimonly*, which is fine-tuned on the WoW dataset to retrieve documents relevant to a query composed only of a response claim, and c) *DPR-WoWft-ctx*, which is also fine-tuned on WoW dataset but uses both the context as well as the response as a query (training details are provided in Appendix C). For DPR-based methods we retrieve the top 100 documents. A document is relevant if it contains a gold evidence sentence.

[5] www.github.com/UKPLab/fever-2018-team-athene

Model	Recall
DPR-original	44.6
DPR-WoWft-claimonly	48.9
DPR-WoWft-ctx	64.1
Wiki-claimonly	60.8
Wiki-ctx	**75.0**

Table 3: Document recall for the test set. Incorporating dialogue context in document improves performance on both WikiAPI and DPR.

	Recall@5	
Model	DPR-WoWft-ctx	Wiki-ctx
Ret-only-claim	67.1	70.1
Ret-with-context	**69.3**	**75.4**

Table 4: Evidence sentence Recall@5 for the test set.

We present the document recall results in Table 3. WikiAPI methods outperform DPR-based methods. Both methods show better performance when dialogue context is used in retrieval. DPR is typically able to retrieve documents with the correct topic but often fails to retrieve a relevant evidence sentence. Entity linking is crucial for fact-checking in dialogue and WikiAPI is able to leverage that capability for better performance.

5.2.2 Evidence Sentence Selection

In evidence sentence selection, a final set of top k evidence sentences are chosen from the set of documents D_c retrieved in the previous step for claim c. First, we create a candidate evidence sentence set S_c by taking the union of all sentences in D_c. We fine-tune a Bert-base model for ranking the candidate sentences in S_c. The model is trained to predict -1 for irrelevant evidence and 1 for relevant evidence for a given claim. We use the context-response pairs from the WoW dataset for training the model. Besides using randomly selected evidence sentences, to create hard negative examples for training, we also chose sentences from the set of articles K_i shown to the wizard during WoW data collection. These sentences are close in content and topic to the gold evidence sentence and form hard negative candidates for the model. At test time, we use the evidence sentences in the top k rank with a score of more than 0. Similar to document retrieval, we created two versions of the model: 1) Ret-with-context, and 2) Ret-only-claim, based on whether the last two utterances of the dialogue context were included in the input to the BERT model. We present the performance of the models in Table 4 for two of the best per-

	Oracle-Evidence		Wiki-Evidence		DPR-Evidence	
Model	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
DNLI	43.3	35.4	39.1	31.5	38.4	29.5
DECODE	37.8	30.3	35.3	25.3	34.5	22.5
VitaminC	54.6	53.7	45.7	43.4	45.6	43.9
CorefBert-Colloquial	61.4	60.0	47.6	45.2	46.4	41.1
Colloquial	63.0	62.2	48.0	46.3	48.7	46.3
Aug-WoW	**68.8**	**68.6**	**51.5**	**51.1**	**51.5**	**50.0**

Table 5: Results for claim verification on the test set. We experiment with three types of evidences and report Accuracy and Macro F1 scores in percentage. Aug-WoW outperforms all baselines across all settings.

forming document retrieval models Wiki-ctx and DPR-WoWft-ctx. We find that recall@5 values for both models are higher when dialogue context is added as an input with the claim.

5.3 Claim Verification

In claim verification, a claim c is classified as SUPPORTED, REFUTED, or NEI given a context C and evidence sentences set S_c.

5.3.1 Baselines

DNLI (Welleck et al., 2019) Dialogue NLI dataset contains sentence pairs labeled as entailment, neutral, or contradiction derived from dialogues. Entailment maps to SUPPORTED, neutral maps to NEI, and contradiction maps to REFUTED in our task. We train a Bert-base model on their training set of 310,110 data points.

DECODE (Nie et al., 2021) Dialogue Contradiction Detection dataset contains both human-human and human-bot contradictory dialogues. The train set contains 27,948 data points with two labels contradiction and non-contradiction. We train a Bert-base model with the last two utterances of the context and the response as input to the model.

VitaminC (Schuster et al., 2021) VitaminC is a large-scale fact verification dataset which is based on contrastive claim-evidence pairs created from Wikipedia edits. They train models that avoid claim-only biases and are more sensitive to changes in the evidence. We use their ALBERT-base model finetuned on FEVER (Thorne et al., 2018) and their VitaminC dataset.

Colloquial (Kim et al., 2021) It contains colloquial claims converted from FEVER dataset claims into colloquial style. It has 410k colloquial claim-evidence pairs in the training set and is well aligned to our task because of its colloquial nature. We fine-tune a Bert-base model on this dataset.

CorefBert-Colloquial (Ye et al., 2020) is one of the best performing models on FEVER and is designed to better capture and represent the coreference information. We use their model which uses kernel graph attention network (KGAT) (Liu et al., 2020) and fine-tune it on Colloquial claims.

Aug-WoW We propose a novel model which is trained on weakly supervised training data. DIALFACT is meant to be used only for validation and test, and we do not train a model on DIALFACT to avoid creating a model which can simply learn to solve the dataset instead of the task. Instead, we leverage the techniques described in section 4.1.1 to create synthetic training data for each category of claims. For SUPPORTED claims, we use the claim-evidence pair from the original WoW dataset. We use the *Lexical* baseline from section 5.1 to filter out Non-Verifiable claims, which leads to 46,934 SUPPORTED claims. We follow the methods *Negation* and *Substitution* from section 4.1.1 to create 38,895 REFUTED claims. We create NEI claims using two methods: 1) For every context-claim-evidence triplet, we substitute the evidence with random unrelated evidence. 2) We use the *Generation* approach from section 4.1.1 to condition the generation on random evidence. We select a subset of 40,000 NEI claims from the two approaches. We fine-tune the *Colloquial* baseline model on this synthetic dataset. The input to the model is the sequence of the last 2 context utterances separated by [EOT] token, followed by the claim.

For all Bert-based models, all evidence sentences are concatenated together. More details about training the baselines are provided in Appendix C.

5.3.2 Results

Table 5 summarizes the results for claim verification on the test set. NON-VERIFIABLE claims are included in the NEI category. We experiment with three evidence retrieval settings - 1) Oracle Evidence, where we use gold evidence, 2) Wiki-Evidence, where we use Wiki-ctx for document retrieval and Ret-with-context for evidence selection, and 3) DPR-Evidence, where we use DPR-WoWft-ctx for document retrieval and Ret-with-

Context Response Evidence	Biathlon means two sports right? What is the other sport? Biathlon combine the two sports into one event called the cross country ski race. It's a lot of fun! Biathlon: The biathlon is a winter sport that combines cross-country skiing and rifle shooting.	Response type: Generated DNLI: S, CorefBERT-Colloquial: S, DECODE: R, VitaminC: NEI, Colloquial: S, AugWoW: R, Human: R
Context Response Evidence	Do you know if professional cheerleaders make a lot of money? The whole point of cheerleading is to show off their skills, so I'm sure they get paid a lot of money. Cheerleading: Cheerleading originated in the United States with an estimated 1.5 million participants in all-star cheerleading.	Response type: Generated DNLI: S, CorefBERT-Colloquial: NEI, DECODE: R, VitaminC: S, Colloquial: S, AugWoW: NEI, Human: NEI
Context Response Evidence	Japanese is even harder, the language is difficult to speak. The origins of the language lie in the prehistoric times when many cultures spoke to one another. Japanese language: Little is known of the language's prehistory, or when it first appeared in Japan.	Response type: Generated DNLI: S, CorefBERT-Colloquial: NEI, DECODE: S, VitaminC: NEI, Colloquial: NEI, AugWoW: NEI, Human: R
Context Response Evidence	I might recognize if I heard it. Who else did you listen to in the 90s? I also listened to another group Dave Grohl was apart of called Them Crooked Vultures. It was not one of his best groups. Dave Grohl: He is the drummer and co-founder of the rock supergroup Them Crooked Vultures.	Response type: Written DNLI: S, CorefBERT-Colloquial: R, DECODE: R, VitaminC: NEI, Colloquial: R, AugWoW: R, Human: S

Table 6: Sample dialogue contexts, claims, evidences and model predictions. We also indicate whether the response is automatically generated or human written. Here S stands for SUPPORTED and R for REFUTED.

context for evidence selection. We set the maximum evidence to 5. In all three settings, Aug-WoW outperforms baselines and the performance of all baselines drops when retrieved evidence is used compared to when oracle evidence is used. This indicates that evidence retrieval is an important step for this task. Even with oracle evidence, none of the models achieve an accuracy higher than 70%, which leaves abundant opportunity for future improvements. Colloquial baseline is the closest to Aug-WoW since it has been trained on conversation-like colloquial claims. Although Colloquial and CorefBert-Colloquial perform better than VitaminC with oracle evidence, the contrastive nature of VitaminC helps it perform better with retrieved evidences. We report performance on a two-way classification experiment in Appendix B (Table 11) where we combine REFUTED and NEI into a single class named NOT-SUPPORTED.

5.3.3 Discussion

We present sample dialogue contexts, claims, oracle evidence for the claims along with model predictions in Table 6. We found that models tend to incorrectly predict a REFUTED or NEI response as SUPPORTED when there is significant overlap between the evidence and the claim while ignoring the semantics. The first example illustrates this point where the presence of terms "biathlon" and "cross country skiing" misleads some models to predict SUPPORTED incorrectly. Similarly, models predict SUPPORTED or REFUTED for a NEI claim due to word overlap between claim and evidence, as shown in the second example. Models also often fail to perform complex and commonsense-based reasoning during verification. In the third example, although humans can reason that the claim is REFUTED by the evidence, all models fail to correctly classify the claim. Finally, models struggle with lexical biases and separating the colloquial part of a claim from its factual parts. In the fourth example, although there is significant overlap between the claim and the evidence, models are fooled by the presence of the word "not one of", and predict a SUPPORTED claim as REFUTED.

6 Conclusion

We propose a new benchmark, DIALFACT, for fact-checking in dialogue created based on grounded dialogues from the Wizard-of-Wikipedia dataset. Besides human-written response claims, we also create synthetic claims with operations such as contradiction, infilling and substitutions. We hire qualified crowd workers to annotate responses into NONVERIFIABLE, SUPPORTED, REFUTED, or NOTENOUGHINFORMATION categories along with corresponding evidence. We point out empirically that existing fact-checking models trained on non-dialogue data fail to perform well on our task. We demonstrate how to leverage automatically generated responses as weak supervised signals to improve performance. We hope that DIALFACT can facilitate factual consistency modeling and evaluation research in the dialogue community.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

David DeVault and Matthew Stone. 2007. Managing ambiguities across utterances in dialogue. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Roverto, Italy. SEMDIAL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Sarik Ghazarian, Zixi Liu, Akash S M, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2021. Plot-guided adversarial example construction for evaluating open-domain story generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4334–4344, Online. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking. *arXiv preprint arXiv:2108.11896*.

Prakhar Gupta, Yulia Tsvetkov, and Jeffrey Bigham. 2021. Synthesizing adversarial negative responses for robust response ranking and evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3867–3883, Online. Association for Computational Linguistics.

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 *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

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

Byeongchang Kim, Hyunwoo Kim, Seokhee Hong, and Gunhee Kim. 2021. How robust are fact checking systems on colloquial claims? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1535–1548, Online. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval*, pages 639–649, Cham. Springer International Publishing.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China. Association for Computational Linguistics.

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

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Darsh J Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, Hong Kong, China. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, Online. Association for Computational Linguistics.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, Hong Kong, China. Association for Computational Linguistics.

Andrew Trask, Phil Michalak, and John Liu. 2015. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*.

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

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

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Jianshu Chen Wenhu Chen, Hongmin Wang, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact : A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

A Lexical Biases

Following Schuster et al. (2019), we measure the Local Mutual Information (LMI) to measure the correlation between bigrams in the claims (w) and the categories l, defined as follows: $LMI(w,l) = p(w,l)log\left(\frac{p(l/w))}{p(l))}\right)$. We present the top bigrams in REFUTED claims and their LMI value in Table 7. The top bigrams in DIALFACT do not include obvious negations such as "do not", "is not", are mostly topical in nature, and the $p(l/w)$ value is low with the Refute label. Investigating generated and written claims separately, we found that bigrams such as "does not, only one, did not, are not" had higher $p(l/w)$ in written claims compared to generated claims for REFUTED category, although their LMI

All			Labelled			Written		
Bigram	LMI	p(l/w)	Bigram	p(l/w)	p(l/w)	Bigram	p(l/w)	p(l/w)
he was	396	0.45	he was	692	0.40	only one	201	0.95
was born	362	0.64	singer songwriter	471	0.61	referred as	169	0.83
spectrum visible	195	0.80	spectrum visible	447	0.82	drama school	163	0.89
visible light	188	0.76	visible light	431	0.74	harry potter	160	0.60
on spectrum	186	0.73	on spectrum	431	0.78	pins are	158	0.83
an american	177	0.50	an american	391	0.47	only be	152	0.89
singer songwriter	173	0.61	songwriter actor	322	0.67	written by	143	0.77
was released	158	0.53	elvis presley	273	0.47	on visible	138	0.69

Table 7: Top 8 LMI(10^{-6}) ranked bigrams in the test set for REFUTE category.

values were not high. Finally, there is significant overlap between the top bigrams for different categories, suggesting an absence of obvious lexical biases in the dataset.

We perform another experiment where we train Aug-WoW with no evidence included during training and testing. This baseline *Aug-WoW-claimonly* achieves 33.2% accuracy and 28.9% macro F1 score on the DIALFACT test set. Thus, a model can not exploit lexical cues in the claims of DIALFACT to obtain good performance.

B Supplementary Results

We present the claim verification results on the validation set in Table 8. The trend in performance is similar to the trend observed in the test set reported in 5. In our human studies discussed in subsection *Data validation* of section 4.4, we observe that workers confuse between REFUTED and NEI labels. Furthermore, there are cases where the workers can miss finding an evidence which refutes a claim on Wikipedia and label the claim as NEI even though they are instructed to find and verify a claim by visiting Wikipedia. Similar findings were reported in other fact-checking tasks (Jiang et al., 2020). Hence we perform another experiment where we combine REFUTED and NEI into a single class, and name it NOT-SUPPORTED. We present the claim verification results on test set for this setting in Table 11. The performance of all baselines is higher since the task is transformed to a 2-way classification task from a 3-way classification task. Aug-WoW performs the best in this setting.

In Section5.3.2, we discuss results where NON-VERIFIABLE claims are included in the NEI category. In Table 10, we present the results for 3-way classification on test set where NON-VERIFIABLE claims with NEI-PERSONAL labels are removed, that is, only Verifiable claims are kept for NEI labelled claims. The trends in results are similar to the ones observed in Table 5.

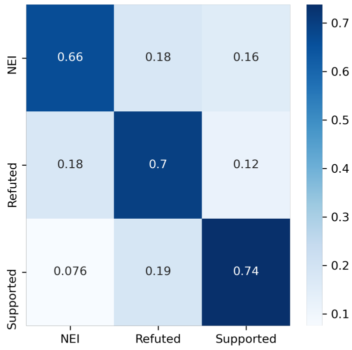

Figure 2: The Confusion matrix of Aug-WoW model.

In Table 9 we present the claim verification results on the Test set using oracle evidence on Generated and Written claims separately. The performance of all models is lower on Generated claims compared to Written claims. This is expected since as we mentioned in “Final claim set creation” in section 4.1.1, the Generated claims were chosen from a larger candidate claims set based on the difficulty of existing models to classify those claims. Thus Generated claims in DIALFACT are more challenging. Furthermore, Aug-WoW’s performance is high on both types of claims, however, the gain in its performance on Written claims is higher on Written claims compared to Generated claims.

In Table 12 we present the claim verification results on the test set with Aug-WoW model ablations. In Aug-WoW-noctx we do not concatenate the dialogue context, and in Aug-WoW-BertLarge we use the Bert-Large model as base architecture. Aug-WoW-noctx is comparable to Aug-WoW, ith slightly lower performance with Oracle evidence. Although Aug-WoW-BertLarge performs better with oracle evidence, it performs poorly with retrieved evidence. This indicates that it is more sensitive to the evidence quality.

We show the confusion matrix of our Aug-WoW model in Figure 2. Aug-WoW has the lowest per-

	Oracle-Evidence		Wiki-Evidence		DPR-Evidence	
Model	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
DNLI	42.0	34.9	39.0	31.1	38.2	30.1
DECODE	31.6	29.2	33.5	25.7	31.1	21.2
VitaminC	64.4	58.8	48.4	44.0	48.5	44.8
CorefBert-Colloquial	64.5	63.0	46.8	44.4	46.2	42.4
Colloquial	68.4	64.2	50.0	47.4	51.9	48.8
Aug-WoW	**71.3**	**71.3**	**51.4**	**51.2**	**50.6**	**49.9**

Table 8: Results for claim verification on the validation set. We experiment with three types of evidences and report Accuracy and Macro F1 scores in percentage. Aug-WoW outperforms all baselines across all settings.

	Generated		Written	
Model	Accuracy	Macro F1	Accuracy	Macro F1
DNLI	50.9	38.4	34.8	31.0
DECODE	36.5	30.4	39.3	30.1
VitaminC	48.9	42.1	60.8	60.3
CorefBert-Colloquial	56.9	51.6	66.4	65.5
Colloquial	61.3	56.9	64.7	64.6
Aug-WoW	**63.9**	**60.7**	**74.2**	**74.0**

Table 9: Results for claim verification on the test set for Generated and Written claims.

formance on NEI claims and highest confusion between NEI and Refuted classes.

C Implementation Details

First we discuss the implementation details for claim generation techniques in section 4.1.1. For Negation we use the implementation from fever-2 baseline[6] (Thorne et al., 2019). For the T5 model in *Mask-and-Fill* and Blenderbot model in *Generation* approach, we use the models and training scripts available in the Hugging Face's Transformers repository[7]. Blenderbot was finetuned on full WoW training dataset with batch size of 40.

We next discuss the implementation details for the document retrieval methods. For WikiAPI method, Kim et al. (2021) pointed out that WikiAPI method naively retrieves documents related to filler words such as "I", "Yes", "They" etc. frequently. In our implementation of WikiAPI we mitigate this issue by filtering out such colloquial phrases by using a manually created stopwords list. We remove the stopwords from the candidate set of entities on which MediaWiki API is called. Our experiments showed significant improvement in the quality of the returned documents. For DPR, we use the *wiki_dpr* dataset available in the Hugging Face Datasets library[8] for document retrieval. It contains 21M passages from wikipedia along with their DPR embeddings. The wikipedia articles are split into multiple, disjoint text blocks of 100 words as passages. We retrieve top 100 documents per claim. We finetune the claim encoders for *DPR-WoWft-claimonly* and *DPR-WoWft-ctx* using the original DPR implementation[9]. The original biencoder was trained on natural questions dataset. We only fine-tune the question encoder of the DPR model. DPR training data consists of positive, random negatives and hard negative pairs. For positive claim-evidence document pairs, we use the response-knowledge sentence pairs in the original WoW dataset, where we filter out NON-VERIFIABLE claims using the *Lexical* baseline from section 5.1. For hard negatives, we follow the instructions in the DPR repository and mine hard negatives using the original DPR index and encoder (facebook/dpr-question_encoder-single-nq-base) itself. Specifically, we use DPR to retrieve top 2 evidences per claim and use them as a hard negative if they are not the same as the original knowledge sentence for the claim in the WoW dataset. We finetune the base DPR encoder on the aforementioned constructed data and convert only the question encoder checkpoints into Hugging Face model format.

We next discuss the implementation details for the models for claim verification 5.3. For VitaminC, we use the tals/albert-base-vitaminc-fever model available in their repo[10]. We finetune CorefBERT-base for CorefBERT and use the official code from the authors[11]. We train AugWoW and Colloquial models using the code from the VitaminC repo[12] on a machine with 4 NVIDIA A100 GPUs and train batch size of 100. We use the validation set performance for model selection.

[6] www.github.com/j6mes/fever2-baseline

[7] www.github.com/huggingface/transformers/

[8] www.huggingface.co/datasets/wiki_dpr

[9] www.github.com/facebookresearch/DPR

[10] www.github.com/TalSchuster/VitaminC

[11] www.github.com/thunlp/CorefBERT/tree/master/FEVER

[12] www.github.com/TalSchuster/VitaminC

	Oracle-Evidence		Wiki-Evidence		DPR-Evidence	
Model	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
DNLI	43.8	33.7	41.3	32.2	41.3	30.4
DECODE	41.8	31.7	39.0	26.7	38.1	23.8
VitaminC	50.4	50.8	40.6	40.3	40.8	40.7
CorefBert-Colloquial	64.1	61.9	50.1	46.5	50.0	43.0
Colloquial	62.7	61.4	48.0	45.9	49.6	46.3
Aug-WoW	**68.8**	**68.0**	**51.4**	**50.1**	**52.6**	**49.3**

Table 10: Results for claim verification on the test set for 3-way classification where Non-Verifiable claims with NEI-Personal labels are removed and for NEI only Verifiable claims are kept. We report Accuracy and Macro F1 scores in percentage.

	Oracle-Evidence		Wiki-Evidence		DPR-Evidence	
Model	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
DNLI	48.1	46.5	47.2	46.3	43.9	42.0
DECODE	65.4	62.5	63.2	52.2	62.3	47.1
VitaminC	74.5	66.3	70.2	60.6	69.3	61.1
CorefBert-Colloquial	72.3	71.8	63.3	62.9	57.7	57.7
Colloquial	76.6	75.3	66.4	65.1	63.5	63.0
Aug-WoW	**80.6**	**78.8**	**69.0**	**67.4**	**68.2**	**67.3**

Table 11: Results for claim verification on the test set for 2-way classification - SUPPORTED and NOT-SUPPORTED. We combine REFUTED and NEI into NOT-SUPPORTED. We report Accuracy and Macro F1 scores in percentage.

D AMT Instructions

We present the screenshot of the annotation interface is shown in Figure 3. Workers were paid an avergae of $8-10 per hour across all tasks. For the claim labelling task, workers were told that they will be shown a conversation between two speakers, some previously created responses to the conversation, and some Wikipedia knowledge snippets related to the response (which we will call evidence henceforth). They will label some dialogue responses which could belong to one of the 3 categories mentioned below.

Supported: The response should exclusively use factual information which can be verified by the given evidence sentences and is correct or true in light of the evidence. A response is verifiable if evidence could be retrieved from Wikipedia, which decreases the uncertainty about the truthfulness (or falsehood) of the statement.

Example 1:

- Context: I think Jazz is an American creation!
- Evidence: Jazz has roots in West African cultural and musical expression, and in African-American music traditions including blues and ragtime, as well as European military band music.
- Response: Its roots include African-American music traditions including blues and ragtime
- Explanation: Response is natural and can be verified from the evidence.

Example 2:

- Context: What are the three different waterfalls Niagra is made from? Can you please share with me?
- Evidence: From largest to smallest, the three waterfalls are the Horseshoe Falls, the American Falls, and the Bridal Veil Falls.
- Response: The three waterfalls are the Horseshoe Falls, the American Falls and the Bridal Veil Falls.
- Explanation: Response is natural and can be verified from the evidence as all facts mentioned are correct.

Refuted: The response contains factual information which is “incorrect” or “false” in light of the evidence, that is it contradicts the evidence. The response should be marked refuted if even a small part of the response is incorrect.

Example 1:

- Context: I think Jazz is an American creation!
- Evidence: Jazz has roots in West African cultural and musical expression, and in African-American music traditions including blues and ragtime, as well as European military band music.

	Oracle-Evidence		Wiki-Evidence		DPR-Evidence	
Model	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Aug-WoW-noctx	68.1	68.1	**52.4**	**52.3**	**52.4**	**51.3**
Aug-WoW-BertLarge	**70.9**	**70.9**	45.8	44.6	43.5	39.1
Aug-WoW	68.8	68.6	51.5	51.1	51.5	50.0

Table 12: Results for claim verification on the test set with Aug-WoW model ablations.

Context	A: I prefer to eat fish that is not farm raised due to the pesticides in the food. B: Yes the two most common are atlanticcod and pacific cod A: Most cod sold in stores is farm raised, and also the cod you eat in restaurants.
Responses (Generated)	Response 1: There are other varieties of cod as well, like the black, red, white, and yellow Evidence: Cod flesh is moist and flaky when cooked and is white in colour. It change colour at certain water depths. It has two distinct colour phases: gray-green and reddish brown Labels: Factual, Refuted Response 2: I read that it is a popular food with a mild flavor and a dense flaky flesh Evidence: Cod is popular as a food with a mild flavour and a dense, flaky white flesh. Labels: Factual, Supported Response 3: I read that it is a large fruit with a bunch flower and a fleshy petals. Evidence: Cod is popular as a food with a mild flavour and a dense, flaky, white flesh. Labels: Factual, Refuted
Context	A: Elvis's first RCA single was "Heartbreak Hotel" released in 1956 and became a number one hit in US. B: Right, he became popular pretty quickly! When did he die?
Responses (Written)	Response 1: Some think he died August 16, 1977. He helped pioneer the popular sound of rock and roll. Evidence: Elvis Aaron Presley (January 8, 1935 – August 16, 1977) was an American singer, musician, and actor. He became the leading figure of the newly popular sound of rock and roll. Labels: Factual, Supported Response 2: Some think he died August 25, 1988. He helped pioneer the popular sound of rap music. Evidence: Elvis Aaron Presley (January 8, 1935 – August 16, 1977) was an American singer, musician, and actor. He became the leading figure of the newly popular sound of rock and roll. Labels: Factual, Refuted Response 3: I am trying to remember when he died. But most people in Russia see him as an idol. Evidence: Elvis Presley - He became the leading figure of the newly popular sound of rock and roll. Labels: Factual, NEI

Table 13: We present two examples from DialFact dataset: The top context has responses which were automatically generated and then labelled. The bottom context has responses written and then labelled. The labels and evidence are shown below the responses.

- Response: Its roots include American music traditions including blues and ragtime
- Explanation: Roots are African-American, not American.

Example 2:

- Context: What are the three different waterfalls Niagra is made from? Can you please share with me?
- Evidence: From largest to smallest, the three waterfalls are the Horseshoe Falls, the American Falls and the Bridal Veil Falls.
- Response: The three waterfalls are the Horseshoe Falls, the American Falls and the Sommer Falls.
- Explanation: One of the falls is incorrect based on the evidence.

Not Enough Information: The response can not be verified (supported or refuted) with Wikipedia evidence. Moreover, for this response, it is allowed to use information/knowledge that might not be available in Wikipedia but you assume to be general knowledge, e.g. that 90s refers to the time span from 1990 to 1999.

Example 1:

- Context: I think Jazz is an American creation!
- Evidence: Jazz has roots in West African cultural and musical expression, and in African-American music traditions including blues and ragtime, as well as European military band music.
- Response: Jazz is now played in all parts of the world except Russia.

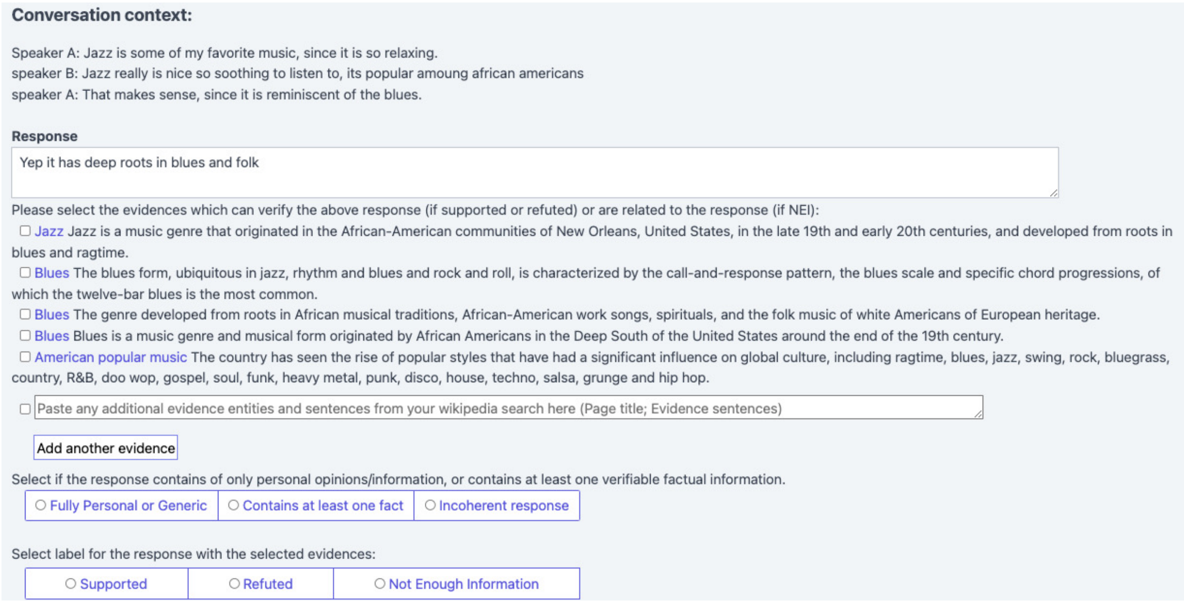

Figure 3: Annotation interface for claim labeling. Workers are shown a conversation context, a claim or response to the context, and evidence sentences from Wikipedia related to the response. They are asked to add any additional evidence necessary for labelling. They first select if the response is VERIFIABLE or NON-VERIFIABLE. Then they select one of the categories - SUPPORTED, REFUTED AND NOT ENOUGH INFORMATION.

- Explanation: The response is not a personal opinion and the provided evidence can't be used to verify the stated fact.

Example 2:

- Context: What are the three different waterfalls Niagra is made from? Can you please share with me?
- Evidence: From largest to smallest, the three waterfalls are the Horseshoe Falls, the American Falls and the Bridal Veil Falls.
- Response: I think three waterfalls all intersect multiple times. I am trying to remember the names.
- Explanation: The stated fact can not be verified from the evidence.

We ask workers to do the following:

- Read the context carefully and if writing or editing a response, write minimum of 9 words.
- The label should be exclusively based on the response and the selected evidence sentences.

We ask workers to NOT do the following:

- While writing or editing a response please avoid typos and mis-spelling as much as possible.
- While writing or editing a response, do not use “know-it-all” phrases such as "did you know" in your responses - e.g., the response "did you know that the Berlin Wall was demolished in 1989" will not be accepted.

Personal/generic response: We give workers some examples of personal response. The response should not make any factual claim that could be verified using Wikipedia or any knowledge source. It can contain facts that are personal opinions or background of the speaker, but no fact pertinent to encyclopedic knowledge. The response should be a good follow-up to the conversation.
Example 1:

- Context: I do not understand why some people enjoy hunting.
- Evidence: Hunting is the practice of killing or trapping animals.
- Response 1: I enjoy going out in the woods to hunt animals.
- Response 2: Wow interesting. I have mostly used hunting as a means of pest control.
- Explanation: Even if hunting can be used as pest control, it is a personal detail or opinion here.

Example 2:

- Context: It would be perfect to have a family member involved in choosing foster care.

- Evidence: Usually children are taken care of by their parents, legal guardians or siblings.

- Response: Very true, that is why I think it is best when parents or or legal guardians take care of their children, because they are they only ones that love the children.

- Explanation: Although part of the response is present in the evidence, this is a subjective opinion of the speaker.

To start the final task, we ask workers to read the dialogue, the corresponding responses, and the Wikipedia knowledge provided (links and pieces of evidence).

- For each provided response, mark them as SUPPORTED, REFUTED, or NOT ENOUGH INFORMATION.

- if the response consists of only personal opinions or personal information with no verifiable factual information, please mark the corresponding checkbox.

- Please read the instructions and examples in the link above carefully.

- If you select the SUPPORTED or REFUTED option, you must click at least one checkbox as evidence or copy-and-paste sentences from Wikipedia links.

- For NEI, you would generally need to verify the facts in the responses by visiting and searching Wikipedia pages and pasting any related evidence.

- Please edit and correct the responses if they contain any grammatical or spelling mistakes.