Automated Fact-Checking in Dialogue: Are Specialized Models Needed?

Eric Chamoun[1], Marzieh Saeidi*[2] , Andreas Vlachos[1]
[1]Department of Computer Science, University of Cambridge
[2]Synthesia, London
{ec806,av308}@cam.ac.uk, marzieh.saeidi@googlemail.com

Abstract

Prior research has shown that typical fact-checking models for stand-alone claims struggle with claims made in dialogues. As a solution, fine-tuning these models on labelled dialogue data has been proposed. However, creating separate models for each use case is impractical, and we show that fine-tuning models for dialogue results in poor performance on typical fact-checking. To overcome this challenge, we present techniques that allow us to use the same models for both dialogue and typical fact-checking. These mainly focus on retrieval adaptation and transforming conversational inputs so that they can be accurately predicted by models trained on stand-alone claims. We demonstrate that a typical fact-checking model incorporating these techniques is competitive with state-of-the-art models fine-tuned for dialogue, while maintaining its accuracy on stand-alone claims.

1 Introduction

The need for fact-checking is ever-growing as the volume of false claims on social media platforms rises, inspiring researchers to develop automated tools to combat misinformation (Zeng et al., 2021; Guo et al., 2022). Despite the application of automated fact-checking to various use cases, most studies still focus on stand-alone, well-formed claims similar to those found in formal sources like encyclopedias. However, such claims are different from those found in conversations, which often feature incomplete utterances that reference previously mentioned entities or even omit them (Tseng et al., 2021; Varshney et al., 2022). Additionally, conversational utterances often include filler words and casual comments, which complicate the task.

Recently, Gupta et al. (2022) presented DialFact, a dataset for automated fact-checking in dialogue. Their experiments showed that state-of-the-art models, trained on stand-alone well-formed claims, do not perform well on DialFact. To address this, they propose instead to fine-tune models on conversational claims within their dialogue contexts.

In this paper, we first demonstrate that although these models obtain strong results on DialFact, they suffer from a significant decrease in accuracy on stand-alone fact-checking, a form of catastrophic forgetting, i.e. the tendency of a model to forget previously learned abilities after learning from new data (French, 1999). Furthermore, we argue that building a separate model for every real-world scenario is not a practical solution, since each model requires ongoing monitoring and maintenance for long-term reliability (Sculley et al., 2015).

For these reasons we introduce methods for adapted evidence retrieval and input transformation without changing the fact-checking model. We first present a claim detection technique to tackle the low density of factual information in dialogue claims (Figure 1, in red). Additionally, we modify document retrieval to handle both the conversational context and the claim, but place more weight on the latter to reduce noisy retrieval results (Figure 1, in blue). Finally, we enhance sentence retrieval by considering not only the relevance of the retrieved sentence to the claim, but also that of the document it is sourced from (Figure 1, in green).

By incorporating the proposed techniques, a typical fact-checking model can match the performance of state-of-the-art models fine-tuned specifically for dialogue on DialFact, while maintaining its accuracy on FEVER (Thorne et al., 2018), a benchmark for stand-alone well-formed claims. In comparison, fine-tuning the same model for dialogue results in a minimum reduction of 12% accuracy on FEVER due to catastrophic forgetting.

* This work was completed while the author was at Meta AI.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16009–16020
December 6-10, 2023

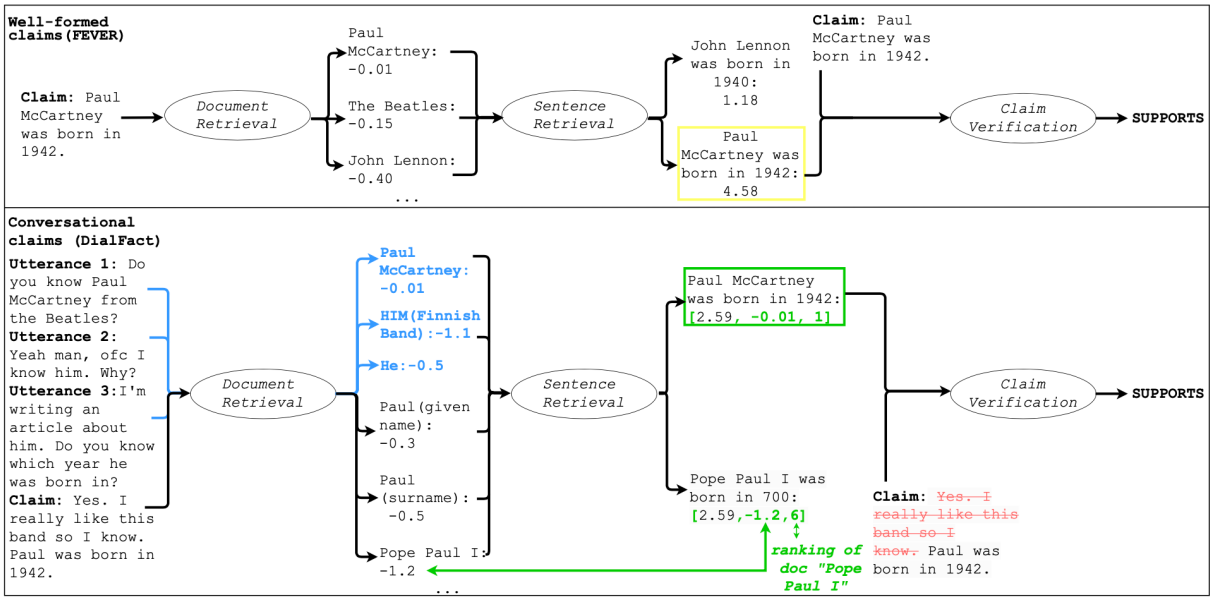

Figure 1: Overview of our approaches for typical and dialogue fact-checking. The proposed techniques are highlighted in blue for document retrieval, green for sentence retrieval and red for claim detection.

2 Fact-checking conversational claims

Fact-checking systems typically consist of three components: a document retriever that returns the most relevant documents to a given claim from a textual source such as Wikipedia, a sentence retriever that selects the most relevant evidence sentence(s) from the retrieved documents, and a claim verification model that classifies the claim as SUPPORTS, REFUTES or NOT ENOUGH INFO (NEI) w.r.t. the evidence. We formulate fact-checking dialogue claims as the task of verifying the last utterance, referred to as the claim c, of a multi-turn conversation $C = \{U_1, ..., U_{n-1}, c\}$. This section presents techniques aimed at improving the dialogue fact-checking performance of a model trained on stand-alone well-formed claims, without requiring any adaptation of the model itself.

Document retrieval In order to take into account the dialogue context of the claim, we return the union of the top k documents that are most similar to the claim, along with the single most relevant document to each utterance in the context:

$$D_c = f(c, k) \cup \sum_{i=1}^{n-1} f(U_i, 1) \qquad (1)$$

, where f is a scoring function (not fine-tuned to dialogues) that takes as input a sentence s and a number k, and returns the top k most relevant documents to s. The proposed method enables capturing the main entities of the conversation, despite the presence of coreference and ellipses. For example, in Figure 1, the name "Paul McCartney" is referred to as simply "Paul" in the conversational claim, making accurate document retrieval difficult. Our approach tackles this challenge without retrieving a large number of irrelevant documents. Indeed, solely retrieving documents related to the claim would involve considering a broad range of entities containing "Paul" in their names. In contrast, the proposed method enables the retrieval of the precise entity under discussion within the conversation, eliminating the need to search through all possible "Paul" entities. Moreover, by focusing on the single most relevant document to each sentence within the context and simultaneously retrieving the top k most relevant documents to the claim, we strike a balance. This approach ensures that the entities in the context are considered, but not to the extent that they overshadow the importance of the claim itself.

Sentence retrieval Just like document retrieval, the performance of sentence retrieval can be greatly impacted by the presence of coreference and ellipses in dialogues, as demonstrated in Figure 1. In the conversational example, the sentence retriever should assign the highest score to the Wikipedia sentence that contains information about the birth year of "Paul". However, in this case, "Paul" could equally refer to either "Paul McCartney" or "Pope Paul I", as documents with these titles were re-

trieved. As a result, the sentence retriever is unable to distinguish the correct evidence and assigns equal scores to sentences from both documents.

To address this issue, we propose incorporating document relevance scores into sentence retrieval. This method capitalizes on the contextual information gathered during document retrieval, making sure it is utilized effectively in sentence retrieval. For instance, as shown in Figure 1, the information that "Paul McCartney" is the most relevant document to the conversation, while "Pope Paul I" is the sixth most relevant, increases the likelihood that the correct evidence is found in the former document.

The proposed technique operates by combining information gathered during document and sentence retrieval as follows:

$$score(s; c) = g(r_s; r_D; R_D) \quad (2)$$

, where g is a parameterized function, s is an evidence sentence in a document $D \in D_c$, r_s is the similarity score of s to the claim c, r_D is the similarity score of D to c, and R_D is the ranking of D among D_c. To train function g, the document and sentence retrieval models are first applied to the DialFact training set examples to generate triples $(r_s; r_D; R_D)$ for each sentence s with respect to a claim. Then, a logistic regression model is trained using these triples as inputs and Boolean values indicating whether each evidence sentence is a piece of evidence according to the gold standard.

Claim Detection Typical fact-checking models are trained on claims that are single-factoid, self-contained sentences, such as those in FEVER (Thorne et al., 2018). However, claims in dialogues often span multiple sentences, and may contain content that is not possible to check, such as "Yes. I really like this band so I know." in Figure 1. This type of information can be challenging for these models to distinguish from the verifiable portion of the claim. To address this issue, we present a technique for identifying the factual information in dialogue claims. It operates by selecting the part of the utterance that has the highest semantic textual similarity to the retrieved evidence. The process begins by splitting the claim into sub-sentences. Next, we use a sentence encoder to generate encodings for each sub-sentence and each retrieved evidence sentence. Finally, the claim is replaced with the sub-sentence that has the highest cosine similarity with the retrieved evidence.

3 Results

Implementation details FEVER (Thorne et al., 2018) is a benchmark dataset comprising well-formed claims derived from Wikipedia. Our approach builds on the state of the art. It employs GENRE[1] (Cao et al., 2021) as our scoring function f for document retrieval following Thorne (2022). For sentence retrieval and claim verification, we leverage the Bigbird-based (Zaheer et al., 2020) and DeBERTa (He et al., 2020) models[2] respectively, from Stammbach (2021). For the evidence enhancement ensemble, we train a logistic regression model g using the scikit-learn library[3]. For claim detection, we leverage SRoBERTa[4] and Spacy's Sentencizer[5] to perform sentence encoding and sentence splitting, respectively. Finally, we used SpanBERT (Joshi et al., 2020) for coreference resolution, StyleFormer[6] for style transfer, and the GPT-2-based model proposed by Tseng et al. (2021) for claim rewrite.

	Document Retrieval	
	Recall	Recall (No NEI)
Claim-only	56.85	60.02
Resolved Claim-only	67.0	72.0
Concatenated Claim + Context (Gupta et al., 2022)	76.5	79.3
Proposed Method	**81.90**	**83.76**

Table 1: Document recall for claim-only and claim+context approaches using GENRE.

Document retrieval results Table 1 presents a summary of the document retrieval results achieved on the DialFact test set. We conduct a comparative analysis of various methods: applying GENRE directly to the claim, applying GENRE to the claim after performing coreference resolution from the context using SpanBERT, employing the approach suggested by Gupta et al. (2022), which entails concatenating the claim with the last two utterances of context and applying the scoring function f to the

[1] https://github.com/facebookresearch/GENRE

[2] https://github.com/dominiksinsaarland/document-level-FEVER

[3] https://scikit-learn.org/stable/index.html

[4] https://github.com/UKPLab/sentence-transformers

[5] https://spacy.io/api/sentencizer/

[6] https://github.com/PrithivirajDamodaran/Styleformer

	FEVER	DialFact
	Accuracy	
FEVER	**79.80**	50.75 (-12.88)
+VitC	76.99 (-2.81)	57.84 (-5.79)
+DialFact	67.03 (-12.77)	61.08 (-2.55)
+Colloquial	65.07 (-14.73)	60.12 (-3.51)
+VitC+DialFact	64.88 (-14.92)	61.99 (-1.64)
+VitC+Colloquial	64.56 (-15.24)	61.10 (-2.53)
+retrieval	**79.80**	51.78 (-11.85)
+VitC+retrieval	76.99 (-2.81)	58.36 (-5.27)
+claimdet	**79.80**	53.30 (-10.33)
+VitC+claimdet	76.99 (-2.81)	59.72 (-3.91)
+retrieval+claimdet	**79.80**	54.56 (-9.07)
+VitC+retrieval+claimdet	76.99 (-2.81)	60.72 (-2.91)
+DialFact+retrieval	67.03 (-12.77)	62.83 (-0.80)
+Colloquial+retrieval	65.07 (-14.73)	60.93 (-2.70)
+VitC+DialFact+retrieval	64.88 (-14.92)	**63.63**
+VitC+Colloquial+retrieval	64.56 (-15.24)	61.54 (-2.09)
AugWoW (Gupta et al., 2022)	60.98 (-18.82)	51.60 (-12.03)
AugWoW+retrieval	60.98 (-18.82)	54.38 (-9.25)

Table 2: Performance analysis on FEVER DEV and DialFact TEST of typical fact-checking models combining our methods versus specialized models for dialogues. The models proposed in this paper are in italic. Best performance per dataset is in bold.

result, and our proposed approach (Section 2). By looking at the results, it is clear that directly using the context substantially improves document recall. This result is expected, as the main entities of a conversation are often repeated numerous times in the context. Additionally, these methods do not depend on the coreference resolution accuracy. The scores also show that our proposed method improves upon that presented in Gupta et al. (2022) by more than 5 percentage points in terms of document recall when $k = 10$ (we select the top 10 documents using the claim). To ensure this performance increase is not merely due to an increase in the average number of retrieved documents, we additionally tested our method with $k = 5$. The document recall, in this case, is equal to 80.07% with an average number of retrieved documents of 7.79. This average includes 2.79 documents retrieved from the context, in addition to the top 5 documents most relevant to the claim. In contrast, other methods retrieve a minimum of 10 documents. The effectiveness of this approach compared to that presented in Gupta et al. (2022) can be explained by the higher emphasis on the claim. Our method focuses on retrieving the single most relevant document to each sentence within the context while simultaneously retrieving the top k most relevant documents to the claim. In contrast, the approach in Gupta et al. (2022) directly applies the model to the concatenation of the claim and context, often resulting in the retrieval of noisy documents.

Final results Table 2 summarizes the claim verification results on the test set. The first group consists of typical fact-checking models trained on FEVER (*FEVER*), fine-tuned using VitaminC (Gupta et al., 2022, *FEVER+VitC*). VitaminC (Schuster et al., 2021) is a large-scale dataset containing examples that are *contrastive*: evidence pairs are almost identical in language and content, except that one supports and the other refutes a claim. Training a fact-checking model on VitaminC has been shown to improve a classifier's sensitivity to subtle changes in evidence. In our case, fine-tuning on VitaminC improves the DialFact accuracy by over 7%, while only causing a decrease of less than 3% on FEVER.

In the second group, we fine-tune the typical models on additional dialogue data from DialFact and Colloquial (Kim et al., 2021), as proposed by Gupta et al. (2022). Specializing the models for dialogue leads to significant improvements in DialFact accuracy. However, this approach results in a substantial loss of up to 15% in accuracy on FEVER due to catastrophic forgetting.

The third group uses the typical models with our proposed enhanced evidence selection and claim detection techniques. Applying these together leads to substantial performance improvements of up to 4% on DialFact, while maintaining the accuracy on FEVER. This is reflected in the DialFact accuracy of *FEVER+VitC+retrieval+claimdet*, which performs similarly to the top specialized models for dialogue fact-checking in group 2 while outperforming them by over 12% on FEVER.

The next group demonstrates the advantages of incorporating our evidence enhancement technique to models fine-tuned for dialogue. Specifically, *FEVER+VitC+DialFact+retrieval* achieves state-of-the-art performance on DialFact, outperforming the best previously published results (Gupta et al., 2022) by 12%.

Finally, in the last group, we compare the DialFact accuracy of *AugWow*, the top-performing model from Gupta et al. (2022), when applied using evidence from the best specialized pipeline in their work, versus using our retrieved evidence. The results show an improvement of slightly less than 3% when employing our retrieved evidence without using any specialized models for dialogue.

Figure 2 illustrates the tradeoff between accuracy on dialogue fact-checking and catastrophic forgetting on FEVER. *FEVER+VitC+retrieval+claimdet*,

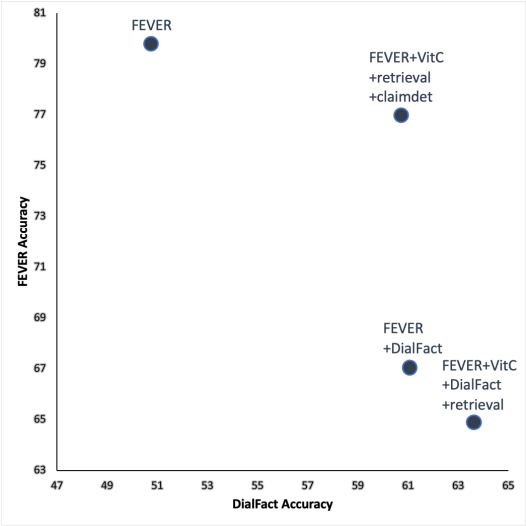

Figure 2: Tradeoff analysis between the accuracy scores on FEVER and DialFact for each model.

a model that combines our proposed methods, optimizes this balance better than the other approaches. It achieves near state-of-the-art results on both FEVER and DialFact, as seen by its placement in the upper right quadrant of the graph. In contrast, the state-of-the-art models for FEVER and DialFact are closer to one of the axes, reflecting their inferior performances on conversational claims and stand-alone formal claims respectively.

4 Qualitative analysis

We conducted a qualitative analysis to assess the advantages and limitations of different approaches.

We initially analyzed the performance improvements offered by our proposed techniques when incorporated to typical fact-checking models by focusing on cases where the central entities in the claim were referred to using pronouns. Our findings showed that most of the times, the document retrieval technique we proposed was still able to successfully identify the appropriate document by taking into account the context (Example 1 in Table B.1). In our study of claims with coreference, we encountered multiple situations where document retrieval was successful but returned multiple documents with similar potential evidence sentences (e.g., birth years of "Pope Paul I" and "Paul McCartney", Figure 1). Our proposed sentence retrieval enhancement technique played a crucial role in these cases by using the context gathered during document retrieval to select the right evidence, resulting in more accurate predictions (Example 2 in Table B.1). We also evaluated instances where the claim contained limited factual information. Frequently, our claim detection method effectively filtered out irrelevant sentences and allowed only the essential information to be processed by the model, leading to accurate predictions (Example 3 in Table B.1).

Additionally, we studied instances where specialized models for dialogue fact-checking outperformed the typical fact-checking model combining our techniques. We found that a common challenge in conversational claims that our approach does not address is indirect claims, such as those made in the form of a question. This limitation is due to the fact that typical fact-checking models are not trained to recognize indirect claims, and our proposed claim detection technique does not resolve this issue (Example 4 in Table B.1).

Finally, we examined cases where *FEVER+VitC+retrieval+claimdet* outperforms specialized models for dialogue fact-checking. We discovered that in instances where a conversational claim could be easily transformed into a well-formed claim through methods such as claim detection, typical fact-checking models often outperformed those specifically designed for dialogue fact-checking. This is because the latter models experience catastrophic forgetting (Example 5 in Table B.1).

5 Conclusion

This paper highlights the significant catastrophic forgetting effects that result from adapting a typical fact-checking model for dialogue. We argue that using separate models for each task is not practical due to the ongoing maintenance cost attached to each. Instead, we propose techniques that allow us to use the same model for both use cases. These mainly focus on retrieval and input adaptation. The model combining these techniques performs comparably to the top specialized models on DialFact while substantially outperforming them on FEVER. We discuss the limitations and societal impact of our approach in the Model Card (Figure A.1).

Limitations

In this study, we present a model designed to autonomously verify claims extracted from dialogues. While our model demonstrates high accuracy on this specific task, it's important to acknowledge its limitations in real-world applications. Our testing benchmark, DialFact, consists of both human-

generated and artificially constructed claims focused on a specific domain, utilizing Wikipedia as a knowledge base. However, this dataset's scope is confined to certain domains, which do not encompass all possible scenarios.

It's worth noting that despite the best efforts of Gupta et al. (2022) in ensuring the high quality of the annotation of DialFact, potential biases and inaccuracies could still exist as in all datasets. Additionally, our model's efficacy has been showcased solely on one dataset. As this task evolves and new datasets emerge, there's a necessity to evaluate our model's performance on diverse datasets to ensure its applicability across a range of scenarios.

Acknowledgements

Eric Chamoun is supported by an EPSRC-funded studentship. Andreas Vlachos is supported by the ERC grant AVeriTeC (GA 865958).

References

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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.

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.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Dominik Stammbach. 2021. Evidence selection as a token-level prediction task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 14–20, Dominican Republic. Association for Computational Linguistics.

James Thorne. 2022. Evidence-based verification and correction of textual claims. Technical Report UCAM-CL-TR-968, University of Cambridge, Computer Laboratory.

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.

Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin Li, and Hong Yu. 2021. CREAD: Combined resolution of ellipses and anaphora in dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3390–3406, Online. Association for Computational Linguistics.

Deeksha Varshney, Akshara Prabhakar, and Asif Ekbal. 2022. Commonsense and named entity aware knowledge grounded dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1335, Seattle, United States. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *CoRR*, abs/2109.11427.

A Model Card

Model card – Automated fact-checking

Model Details

- Document retrieval:
  - GENRE was developed by researchers at University of Amsterdam, Meta, ENS-PSL University, Inria and University College London, 2021, v1.
  - Autoregressive generative model
  - GENRE addresses the document retrieval task using a sequence-to-sequence architecture to generate the title of the relevant Wikipedia page to a claim.
- Evidence selection:
  - Developed by a researcher at ETH Zurich, 2021, v1.
  - Transformer model
  - The Bigbird-based model is used to predict a similarity score between a claim and a sentence in a document for evidence selection.
- Claim verification:
  - DeBERTa was developed by researchers at Microsoft, 2021, v2.
  - Transformer model
  - When used for natural language inference, DeBERTa predicts whether one or multiple pieces of evidence entail a claim.

Intended Use

- Intended for fact-checking stand-alone well-formed claims and conversational claims after retrieving appropriate evidence from Wikipedia.
- Not intended for fully automated moderation.

Factors

- Identify and tackle misinformation aimed at all groups as everyone it susceptible to facing false claims. A potentially relevant factor is age as a younger person is likely less experienced and therefore, in theory, more susceptible to believing fake news. Another relevant factor is social media use due to the proliferation of misinformation on these platforms that causes heavy users to be more exposed.

Training Data

- The models are trained on FEVER for typical fact-checking. We also fine-tune one of these on VitaminC. We do not explore training exclusively on VitaminC due to its heavy imbalance in favor of Supports and Refutes cases.
- We fine-tune FEVER-trained models on the training sets of DialFact and Colloquial for dialogue. As these datasets are automatically or synthetically created, we train the models on FEVER before fine-tuning on these, following Gupta et al., (2021).

Evaluation Data

- FEVER was chosen to evaluate the performance of fact-checking models on stand-alone well-formed claims. It is one of the most popular typical fact-checking datasets and serves as a good benchmark. VitaminC was not used for evaluation due to its heavy imbalance in favor of Supports and Refutes cases.
- DialFact was chosen as it is the most suitable to evaluate the performance of fact-checking models on conversational claims.

Metrics

- Document retrieval:
  - Recall, which measures the proportion of relevant documents that were retrieved, out of all the relevant documents.
- Evidence selection:
  - Recall@5, which measures the proportion of relevant evidence sentences in the top 5 retrieved sentences for each example, out of all the relevant evidence sentences.
- Claim verification:
  - Accuracy, which measures the percentage of correctly fact-checked examples in the test set.

Limtations and Recommendations

- We show that the typical fact-checking model that combines our approaches is effective both on conversational and stand-alone formal claims. However, FEVER and DialFact test sets are not comprehensive as they only contain a small set of the possible claims for each use case. Therefore, we do not expect the model to generalize well on real-world data which is unlike the data it was evaluated on, so we advise against it.
- A potential risk would be to use this model and automatically assume its output is true, which would contribute to misinformation in case it is not correct.

Figure A.1: Model card for the fact-checking model presented in this work.

B Qualitative analysis examples

Context	That's awesome. Do you take part in cheerleading competitions?	FEVER+VitC: S FEVER+VitC+retrieval+claimdet: S FEVER+DialFact: S FEVER+DialFact+retrieval: S FEVER+VitC+DialFact: S FEVER+VitC+DialFact+retrieval: S Human: S
Claim	Somtimes its is more of a intense physical activity now-a-days but sometimes we still do chants.	
Enhanced Evidence	Cheerleading can range from chanting slogans to intense physical activity.	
Context	Was Appetite for Destruction from Guns N' Roses received well?	FEVER+VitC: NEI FEVER+VitC+retrieval+claimdet: R FEVER+DialFact: NEI FEVER+DialFact+retrieval: R FEVER+VitC+DialFact: NEI FEVER+VitC+DialFact+retrieval: R Human: R
Claim	Yes it reached number five on the "Billboard" 200 two years after its release.	
Evidence	Within three weeks, "Blank Space" reached number one on the US "Billboard" Hot 100 following "Shake It Off"	
Enhanced Evidence	'Guns N' Roses' debut album, "Appetite for Destruction" (1987), reached number one on the "Billboard" 200 a year after its release.	
Context	I have heard so much about John Grisham. Can you tell me who he is?	FEVER+VitC: NEI FEVER+VitC+retrieval+claimdet: S FEVER+DialFact: S FEVER+DialFact+retrieval: S FEVER+VitC+DialFact: S FEVER+VitC+DialFact+retrieval: S Human: S
Claim	I sure can!, *John Ray Grisham Jr.is an American bestselling writer, attorney, politician, and activist and hes best known for his popular legal thrillers.* You gotta read the, they are awesome.	
Enhanced Evidence	John Ray Grisham Jr. (; born February 8, 1955) is an American novelist, attorney, politician, and activist, best known for his popular legal thrillers.	
Context	I like all disciplines, i am looking forward to 2018 World Championships, witch are held in Qatar	FEVER+VitC: NEI FEVER+VitC+retrieval+claimdet: NEI FEVER+DialFact: S FEVER+DialFact+retrieval: S FEVER+VitC+DialFact: S FEVER+VitC+DialFact+retrieval: S Human: S
Claim	Have you ever tried rhythmic gymnastics, trampolining and tumbling or any other FIG disciplines?	
Enhanced Evidence	Other FIG disciplines include rhythmic gymnastics, trampolining and tumbling, acrobatic gymnastics and aerobic gymnastics.	
Context	Hello, accounting is very fun, I like to work with numbers.	FEVER+VitC: NEI FEVER+VitC+retrieval+claimdet: R FEVER+DialFact: NEI FEVER+DialFact+retrieval: NEI FEVER+VitC+DialFact: NEI FEVER+VitC+DialFact+retrieval: NEI Human: R
Claim	*Accounting is the reading and listening of medical information.* It seems like serious stuff!	
Enhanced Evidence	Accounting or accountancy is the measurement, processing, and communication of financial and non financial information about economic entities such as businesses and corporations.	

Table B.1: Sample DIALFACT TEST instances highlighting the strengths and weaknesses of the top-performing models. S, R and NEI stand for SUPPORTS, REFUTES and NOT ENOUGH INFO, respectively. With the exception of Example 2, which is specifically chosen to demonstrate the advantages of our retrieval enhancement method, we have selected examples where the correct evidence is retrieved with or without the enhancement. To keep the context concise, we have only included the last turn of the conversation preceding the claim in these examples.

The first example demonstrates the efficacy of our document retrieval technique. Despite the claim only referencing "Cheerleading" using a pronoun, the evidence required to verify it is effectively retrieved from the corresponding document, resulting in accurate predictions from all models.

The second example illustrates the advantages of the retrieval enhancement ensemble. Without using it, no evidence sentence related to "Appetite for Destruction" is found among the top 5 predictions. However, with the enhancement method, the crucial piece of evidence from the "Guns N' Roses" document is assigned the highest score. This is reflected in the claim verification results, as models that did not utilize the retrieval enhancement method produced incorrect predictions.

In the third example, all the models but *FEVER+VitC* generate correct predictions. The specialized models are trained to eliminate extraneous information like "I sure can" and concentrate on the verifiable part of the claim (in italics). Additionally, *FEVER+VitC+retrieval+claimdet* leverages claim detection before claim verification. This approach discards the first and last sentences, retaining only the factual information in the claim, which can be effectively processed by the typical

fact-checking model.

The fourth example highlights a typical challenge in conversational claims that is not addressed by our proposed methods. The claim is made indirectly, in the form of a question, which typical fact-checking models are not trained to identify. This form of *disguised* claim is not addressed either by our claim detection method, leading to NOT ENOUGH INFO predictions from *FEVER+VitC* and *FEVER+VitC+retrieval-claimdet*. In contrast, models fine-tuned specifically for dialogue are able to effectively handle this challenge and generate accurate predictions.

Finally, the fifth example demonstrates a scenario where typical fact-checking models outperform models designed for dialogue fact-checking. All models except for *FEVER+VitC+retrieval-claimdet* produce incorrect predictions. *FEVER+VitC* fails to identify the claim's verifiable portion, while specialized models for dialogue fail to verify the claim with respect to the evidence. However, by applying the claim detection method, *FEVER+VitC+retrieval-claimdet* is left with a well-formed formal claim that it verifies correctly. In cases where the conversational claim can be easily converted into a well-formed claim and does not present significant challenges in dialogue, typical fact-checking models can be more effective due to the catastrophic forgetting effects suffered by models fine-tuned for dialogue fact-checking.

C Claim-transformation techniques

Gupta et al. (2022) state that the dialogue domain poses three main challenges for standard fact-checking models: the coreference and ellipsis phenomena, the low density of factual information in claims, and the colloquial language. In response, we proposed claim-transformation techniques to address these challenges directly. Namely, we explored coreference resolution (Joshi et al., 2020) and claim rewrite (Tseng et al., 2021) to obtain self-contained claims that can be understood independent of previous dialogue context. These involve utilizing information from the previous turns to resolve coreference and ellipses. Additionally, we examined the benefits of applying style transfer to tackle the typical model's struggles with colloquial language. As these techniques were less effective than the techniques discussed in this paper, we include their results in the Appendix.

C.1 Coreference resolution

In DialFact, Gupta et al. (2022) choose to incorporate context by feeding models the concatenation of the claim and the last two utterances preceding it. However, this method requires not only specializing the model for the task but also adds significant noise. Nevertheless, context is crucial for claim understanding due to coreference and ellipses. Therefore, we propose directly addressing these issues by performing coreference resolution to obtain self-contained claims.

We first concatenate the whole dialogue context with the claim. Subsequently, the coreference resolution model predicts coreference clusters in the resulting query. Each cluster consists of ((*span start, span end), span tokens*) pairs, with the first pair being the referent and the remaining ones being its references. Subsequently, we use the span boundaries to replace each reference with its referent and obtain self-contained claims.

We present the results below.

	Document Retrieval	**Evidence Selection**		**Claim Verification** (Oracle Evidence)	
	Recall	Verification Accuracy	Recall@5	Accuracy	Macro F1
Untreated	56.85	**54.19**	**44.06**	**58.73**	**56.66**
Resolved	**67.0**	53.86	42.71	58.60	56.58

Table C.1: Impact of coreference resolution on each stage of the fact-checking process on DialFact DEV.

Coreference resolution improves document retrieval. This result is unsurprising as replacing the mentions with their referents allows GENRE to identify the relevant entities and retrieve their documents. However, this technique harms the evidence sentence selection and claim verification performance. This negative impact can be explained by incorrect resolution cases where the reference is linked to the wrong referent. Indeed, evidence sentence selection is sensitive to a reference resolution mistake as it causes the model to select the most similar sentence in the wrong document. In claim verification, the model is slightly affected because a resolution error changes a claim's label with respect to the gold evidence. For each of these stages, the linking mistakes outweigh the advantages of using this method. In contrast, for document retrieval, the chances of accurate retrieval significantly increase if the resolution is correct.

C.2 Claim rewrite

Another approach we explore to obtain self-contained claims that can be understood independent of context is claim rewrite.

We present the results of applying claim rewrite below.

	Document Retrieval	**Evidence Selection**		**Claim Verification** (Oracle Evidence)	
	Recall	Verification Accuracy	Recall@5	Accuracy	Macro F1
Untreated	56.85	**54.19**	**44.06**	**58.73**	**56.66**
Resolved	**59.56**	53.67	42.12	58.53	56.4

Table C.2: Impact of claim rewrite on each stage of the fact-checking process on DialFact DEV.

Table C.2 shows a very similar pattern to the coreference resolution results. Indeed, claim rewrite improves document retrieval but harms evidence sentence selection and claim verification. However, a manual error analysis reveals numerous rewriting errors and an overall low resolution accuracy. This results in a poorer performance on all fact-checking subtasks than applying coreference resolution.

C.3 Style transfer

Spelling and punctuation mistakes, slang words and colloquialisms make it difficult for a model trained on formal claims to capture the intent of a colloquial claim. Another challenge is the presence of filler words, which significantly affects a retriever's ability to return the correct documents, as shown by Kim et al. (2021). Instead of retrieving relevant documents, the models return documents related to these filler words. In response, we explore style transfer in a bid to formalize the claims. The motivation behind this approach is that it would decrease the claims' wordiness and the presence of expressions that the model may find difficult to understand or recognize. Performing style transfer also expands abbreviations and corrects spelling mistakes and capitalization, which can be key to correct retrieval.

We present the results of applying style transfer below.

	Document Retrieval	**Evidence Selection**		**Claim Verification** (Oracle Evidence)	
	Recall	Verification Accuracy	Recall@5	Accuracy	Macro F1
Untreated	**56.85**	**54.19**	**44.06**	**58.73**	**56.66**
Resolved	55.0	53.52	41.81	55.76	52.62

Table C.3: Impact of style transfer on each stage of the fact-checking process on DialFact DEV.

Table C.3 shows that style transfer does not improve the model performance on any fact-checking component despite the high quality of the formalization. We identify two possible reasons to explain the performance dip caused by this technique. First and most importantly, the generation errors that cause a detail to be omitted in the formalized claim or an incorrect reformulation. Indeed, fact-checking is very sensitive to small changes in a claim. The generation of an equivalent claim needs to be semantically perfect to preserve all details. Consider Example C.1. The reformulation of this claim would score highly on most evaluation metrics for language generation. However, the subtle difference between the two claims that lies in the replacement of *thinnest* with *thin* changes the label of the claim with respect to evidence. Indeed, brown hair is thin compared to red hair but thick compared to fair hair in the gold evidence. In contrast, if we keep *thinnest* then the claim is refuted by the evidence.

Example C.1:

Original Claim: Brown is the color of hair that is the thinnest.
Formal Claim: Brown is the hair color that is thin.
Gold Evidence: Its strands are thicker than those of fair hair but not as much as those of red hair.

The second main reason follows from dataset construction. As the claims are created from Wikipedia passages containing the evidence sentences needed to verify the claims, these often use the same words or formulation as the evidences. This often facilitates the job of evidence sentence selection and claim verification. However, formalizing modifies these words, decreasing the similarity of a claim with its gold evidence. Consider Example C.2. Although the two sentences are semantically equivalent, the original one uses the same words as the evidence.

Example C.2:

Original Claim: I wonder if this associates with the fact that darker hair is more common across the

entire world.
Formal Claim: I am unsure if this is related to the widespread prevalence of darker hair in the world.
Evidence: Black hair is the darkest and most common of all human hair colors globally.