

# EXPLAIN, EDIT, GENERATE: Rationale-Sensitive Counterfactual Data Augmentation for Multi-hop Fact Verification

Yingjie Zhu<sup>1\*</sup>, Jiasheng Si<sup>1\*</sup>, Yibo Zhao<sup>1</sup>, Haiyang Zhu<sup>1</sup>, Deyu Zhou<sup>1†</sup>, Yulan He<sup>2,3</sup>

<sup>1</sup>School of Computer Science and Engineering, Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, China

<sup>2</sup> Department of Informatics, King’s College London, UK

<sup>3</sup> The Alan Turing Institute, UK

{yj\_zhu, jasenchn, yibozhao, haiyangzhu, d.zhou}@seu.edu.cn  
yulan.he@kcl.ac.uk

## Abstract

Automatic multi-hop fact verification task has gained significant attention in recent years. Despite impressive results, these well-designed models perform poorly on out-of-domain data. One possible solution is to augment the training data with counterfactuals, which are generated by minimally altering the causal features of the original data. However, current counterfactual data augmentation techniques fail to handle multi-hop fact verification due to their incapability to preserve the complex logical relationships within multiple correlated texts. In this paper, we overcome this limitation by developing a rationale-sensitive method to generate *linguistically diverse* and *label-flipping* counterfactuals while preserving *logical relationships*. In specific, the diverse and fluent counterfactuals are generated via an Explain-Edit-Generate architecture. Moreover, the checking and filtering modules are proposed to regularize the counterfactual data with logical relations and flipped labels. Experimental results show that the proposed approach outperforms the SOTA baselines and can generate linguistically diverse counterfactual data without disrupting their logical relationships<sup>1</sup>.

## 1 Introduction

Multi-hop fact verification task, which discerns the truth from falsehood based on multiple hops of reliable evidence, becomes crucial in countering misinformation and counterfeit news spread on current social media platforms (Vosoughi et al., 2018; Botnevik et al., 2020), especially in some specific domains such as politics (Alhindi et al., 2018; Ostrowski et al., 2021), science (Wadden et al., 2020, 2022) and public health (Kotonya and Toni, 2020; Sarrouti et al., 2021). However, many recent works

often perform poorly under the multitude of distribution shifts due to an over-reliance on spurious correlations between input text and labels (Gururangan et al., 2018; Schuster et al., 2019; Geirhos et al., 2020). It can potentially be addressed by Counterfactual Data Augmentation (CDA), using counterfactual instances generated by perturbing causal features within the input (Khashabi et al., 2020). Several works have revealed that training with counterfactual data enhances the capability of the model to identify causal features and diminish its reliance on spurious correlations between the input text and the label, thus resulting in the improvement in Out-Of-Domain (OOD) generalization (Vig et al., 2020; Eisenstein, 2022).

In this paper, we seek to generate counterfactuals for multi-hop fact verification, instead of exploring the causal bias for a specific model. However, due to the complex logical relationships within the multi-hop input texts, developing such an approach poses some significant challenges. As shown in the first row of Table 1, most CDA methods are designed for NLP tasks without requiring intricate reasoning over the input, such as the sentiment analysis task (Yang et al., 2021; Howard et al., 2022). Their local modification of the causal feature in a single sentence (e.g., “*amazing*” in Table 1  $\Rightarrow$  “*terrible*”) is difficult to constrain the *logical relationships* between different causal features in multiple correlated texts, resulting in unverifiable counterfactuals. Furthermore, the prior attempt, CrossAug (Lee et al., 2021), is primarily designed to generate counterfactuals for single-hop fact verification via consistently editing the causal features in the claim and in the one piece of evidence (e.g., “*over 30 days*” in the second row of Table 1  $\Rightarrow$  “*less than 10 days*”). Nevertheless, its claim-only based generation strategy struggles to preserve the complex logical relationships when faced with multiple hops of evidence, and fails to ensure *label flipping* and *linguistic diversity* in the counterfactuals, which

\*Equal Contribution.

†Corresponding Author.

<sup>1</sup>The code and datasets are available at <https://github.com/AAAndy-Zhu/RACE>


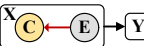
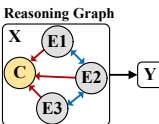
Task	Inference	Input X (label Y)
Sentiment Analysis		This is an <b>amazing</b> book, I'm already immersed in the storyline. ( <b>POSITIVE</b> )
Single-hop Fact Verification		C: Little Miss Sunshine was filmed <b>over 30 days</b> . ( <b>SUPPORTS</b> ) E: Little Miss Sunshine ..., filming began on June and took place <b>over 30 days</b> in Arizona ...
Multi-hop Fact Verification		C: The Ford Fusion was introduced for model year 2006. The Rookie of The Year in the 1997 CART season drives it in the NASCAR Sprint Cup Series. ( <b>SUPPORTS</b> ) E1: <b>Ford Fusion</b> is manufactured and marketed by Ford. <b>Introduced for the 2006 model year</b> , ... E2: <b>Patrick Carpentier</b> competed <b>in the NASCAR Sprint Cup Series, driving the Ford Fusion</b> . E3: <b>The 1997 CART PPG World Series season</b> , ... <b>Rookie of the Year</b> was <b>Patrick Carpentier</b> .

Table 1: Comparison between different tasks.

are crucial for CDA (Joshi and He, 2022).

For multi-hop fact verification, as shown in the third row of Table 1, the set of possible causal features is more complex, and exploring them may necessitate intricate reasoning about the logical relationships between multiple hops of evidence and between the claim and the evidence. For example, the “Patrick Carpentier” in  $E_2$ , which is invisible to the claim, bridges the connection between the causal features “Introduced for the 2006 model year” in  $E_1$  and “Rookie of the Year” in  $E_3$ , thus leading to the alignment of the multi-hop evidence with the claim  $C$  (as shown in the Reasoning Graph). Without considering such complex *logical relationships* within the correlated input, the generated counterfactual claims potentially tend to be unreasonable or unverified. Furthermore, ensuring the *label flipping* and *linguistic diversity* of generated counterfactuals become increasingly difficult with the premise of *logical relationships*, which are critical factors to assure the quality of the counterfactuals.

To address these challenges, we develop a novel pipeline method, RACE (**R**ationale-sensitive **C**ounterfactual **g**eneration), by focusing on the causal features within the rationales extracted from the multi-hop evidence using an explainability method. In specific, for each original instance, the *Explainer* and *Editor* modules are employed to produce the counterfactual evidence that logically corresponds to — but factually distinct from — the original claim. Then, according to the counterfactual evidence, an entity-aware *Generator* generates the counterfactual claims by synthesizing the semantic information across multi-hop evidence. During the above process, the Checking and Filtering modules are used to regularize the reasonableness of the output of each module from different aspects, resulting in fully labeled examples that can be used directly to augment the training data. The

**motivation** here is that these rationales provide the intrinsic semantic and relational information for inferring its label, and present the factual consistency with its claim (Raha et al., 2023).

It should be pointed out that RACE requires no external knowledge as used in Paranjape et al. (2022) besides the original training data, and is able to generate *linguistically diverse* and *label-flipping* counterfactuals while preserving *logical relationships*. Compared to alternative approaches (e.g., ChatGPT (OpenAI, 2022)) (§ 4), training on the counterfactuals generated by RACE reveals the improvement in performance under different settings (§ 5.1), including in-domain, out-of-domain (Paranjape et al., 2022), and challenge settings (Gardner et al., 2020). In addition, the intrinsic evaluation shows that the counterfactual claims generated by RACE are more logical and linguistically diverse than those produced by the baselines (§ 5.3, § 5.4). Finally, we compare the results based on different generation models with baselines, illustrating that our method is generation model-agnostic (§ 5.5).

## 2 Related Works

**Debiasing Fact Verification** A variety of advanced multi-hop fact verification methods have recently emerged in various domains due to the development of pre-trained models (Das et al., 2023). Nevertheless, most models exhibit poor OOD generalization, primarily due to their over-reliance on spurious correlations between inputs and labels (Gururangan et al., 2018; Schuster et al., 2019; Geirhos et al., 2020). Thus, several works focus on the debiasing of fact verification models. Schuster et al. (2019) have identified strong cues for predicting labels solely based on the claim. Zhu et al. (2022) proposed an entity debiasing framework that mitigates entity bias from a cause-effect perspective. Lee et al. (2021) addressed the debiasing of fact verification models by augmenting the data

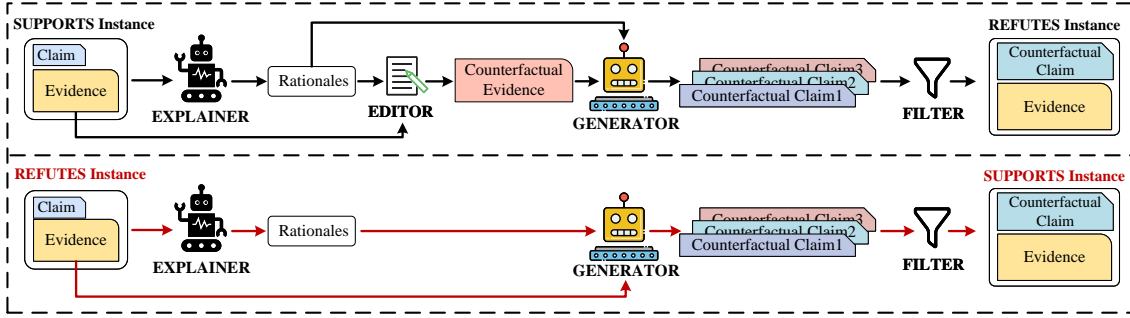


Figure 1: The overall pipeline of RACE. The *SUPPORTS* and *REFUTES* instances are processed differently, as indicated by the black and red arrows, respectively.

with contrastive instances. Atanasova et al. (2022) explored what information is sufficient to verify a claim, and proposed a CDA schema for learning of (in)sufficient information.

**Counterfactual Data Augmentation** There is a growing academic interest in CDA to improve model robustness. Initial studies focus on human-crafted counterfactuals (Kaushik et al., 2020; Gardner et al., 2020). Recently, numerous automatic CDA methods have been proposed for sentiment analysis (Wang and Culotta, 2021; Yang et al., 2021; Howard et al., 2022), question answering (Paranjape et al., 2022; Dixit et al., 2022), and natural language inference (Glockner et al., 2018). However, these methods are primarily targeted to NLP tasks without requiring complex reasoning about the input. Thus, their direct application to the multi-hop fact verification task presents considerable challenges.

### 3 Methodology

Given a claim  $c$  with its associated evidence  $E = (e_1, e_2, \dots, e_n)$ , the aim of multi-hop fact verification is to infer whether the claim is **supported** or **refuted** by the evidence. We denote an instance in the dataset  $D$  as a triplet  $(c, E, y)$ , where  $y \in \{SUP, REF\}$  is the verification label. The goal of RACE is to generate counterfactual data  $(c', E, y')$  or  $(c, E', y')$  that differ in some meaningful way from the original instance  $(c, E, y)$ , where  $y' \neq y$ ,  $c'$  and  $E'$  denote the counterfactual claim and counterfactual evidence, respectively. This setting poses some unique challenges, such as requiring to identify the causal features to be edited, ensuring sound logical relations in evidence editing and claim generation, and avoiding unverifiable claims. Meanwhile, ensuring the semantic diversity and the minimal perturbation of the counterfactuals

can also be challenging. To this end, we propose a general pipeline, RACE, to tackle these challenges.

As shown in Figure 1, our RACE consists of four stages: (I) Explainer: rationale extraction (§3.1), (II) Editor: evidence editing (§3.2), (III) Generator: claim generation (§3.3), (IV) Filtering (§3.4). Note that our method handles *SUP* and *REF* instances differently, as the large difference in generation space between these two types of instances.

#### 3.1 Explainer: Rationale Extraction

Our RACE focuses on identifying the causal features within rationales that can be perturbed. To this end, we use CURE (Si et al., 2023a), a multi-granular rationale extraction method, to simultaneously extract sentence rationales  $R_s$  and token rationales  $R_t$  from the multi-hop evidence  $E$  for both *SUP* and *REF* instances. In essence, the token rationales  $R_t$  reflect the logical correlation within the evidence (blue words in Table 1) and the factual relationship between the claim and the evidence (red words in Table 1). Considering the causal relationship of the rationales to the prediction label (Wu et al., 2022), we regard the extracted rationales as the **causal features** that are to be further processed. The detailed algorithm can be found in Si et al. (2023a).

#### 3.2 Editor: Evidence Editing

In general, entities contained within the multi-hop evidence possess a rich trove of factual knowledge and crucial information (e.g., *date*, *location*, *organization*, *person*, and the correlation between them), facilitating more precise multi-hop fact verification (de Jong et al., 2021; Rani et al., 2023). Therefore, we meticulously design a set of simple entity-based evidence editing rules to control the semantic perturbation while preserving the multi-hop correlation within the evidence, and an Ad-Checking mod-

ule to filter out the under-edited or over-edited evidence. Additionally, Tan et al. (2023) highlight that controlling the generation for *REF* is more challenging due to its significantly broader generation scope compared to *SUP*. As such, we focus on editing the evidence *E* for instances  $(c, E, SUP)$  rather than for instances  $(c, E, REF)$ .

**Editing** We first utilize an off-the-shelf NER tool, Stanza (Qi et al., 2020), to identify various types of **causal entity** *T* from token rationales  $R_t$ . Following Rani et al. (2023), we only retain entities with specific types, including `ORG`, `PERSON`, `DATE`, `GPE`, and `NUM`. Then, we automatically edit the evidence according to the following rules.

- **in-Dataset:** Randomly *replace* entities of type `GPE`, `DATE` and `NUM` with other entities of the same type present in the entire dataset, e.g., *2006 model year*  $\Rightarrow$  *2008 model year* in Table 1.

- **in-Instance:** If all the token rationales in evidence *E* contain two or more `PERSON`/`ORG` entities, their positions are randomly *swapped* between different pieces of evidence, e.g., *Ford* (`PERSON`)  $\Leftrightarrow$  *Patrick Carpentier* (`PERSON`) in Table 1.

- **Consistent Edit:** The same entity token is processed consistently throughout all pieces of evidence, to preserve the multi-hop correlation within the evidence. For example, if an entity is identified in one piece of evidence, it will be consistently replaced or swapped across all pieces of evidence within the instance.

We use the editing rules to produce one edited evidence for each instance based on a random seed. Notably, the `PERSON` and `ORG` entities are unique to each instance, rather than across the entire dataset. Thus, we prefer random in-instance swapping over in-dataset replacing to avoid introducing irrelevant information from the dataset into the edited evidence. See examples in Appendix A.

**Ad-Checking** The random operation in our editing rules may raise the under-editing evidence (i.e.,  $\rightarrow SUP$ ) or the over-editing evidence (i.e.,  $\rightarrow NEI$ ) for *SUP* instances, resulting in the generated claim  $c'$  based on this evidence being an incorrect semantic perturbation compared to its original claim *c*. To this end, we use an existing fact verification model to verify the original claim *c* based on the edited evidence, thus ensuring that this evidence is still valid for further providing to the claim *Generator*. We adopt the RoBERTa (Liu et al., 2019) model, with the concatenation of the edited evidence and

the original claim *c* as input, which is fine-tuned on HOVER (Jiang et al., 2020) dataset with instances labeled as *SUP*, *REF*, and *NEI*. The edited evidence that yields a *REF* prediction is retained as counterfactual evidence  $E'$  (i.e.,  $(c, E') \rightarrow REF$ ). If not, we discard this case for generating counterfactuals. See Appendix B for details.

After Editing and Ad-Checking, we are ready to proceed with the claim generation for the *SUP* and *REF* instances. We retain the original *REF* instance as  $(c, E, T, R_s, REF)$ , and have perturbed the *SUP* instance  $(c, E, T, R_s, SUP)$  to  $(c, E', T', R_s, REF)$ , where *T* and  $T'$  denote the set of original and edited causal entities extracted from the token rationales  $R_t$ , respectively. Up to this step, we generate the counterfactuals  $(c, E', REF)$  by altering the causal entities within the multi-hop evidence for  $(c, E, SUP)$ .

### 3.3 Generator: Claim Generation

As Tan et al. (2023) notes, the direct generation of refuted claims is challenging and may require additional ontology-like mechanisms to ensure that the generation is plausible but reversed. Thus, we opt to generate counterfactual claims  $c'$  that are **supported** by the evidence  $E/E'$  from the instances. Notably, we do not intervene too much in its generation process, apart from regulating the generated claim  $c'$  sensitive to the causal entities  $T/T'$ . This allows us to ensure the linguistically diverse generation while preserving the factual consistency with evidence  $E/E'$ .

**Generation** We first use a pre-trained generation model (e.g., T5 (Raffel et al., 2020)) fine-tuned on the *SUP* instances in FEVEROUS dataset (Aly et al., 2021), using the concatenation of all the gold-standard evidence *E* as input and the corresponding claim *c* as the target text (i.e.,  $E \rightarrow c$ ). Unlike prior work on editing the original claim *c*, this encourages the linguistically diverse generation by synthesizing the semantic and correlation information between the multi-hop evidence.

Then, to ensure that the generated claim  $c'$  presents factual consistency with the evidence  $E/E'$ , we apply constrained beam search decoding (Anderson et al., 2017; Post and Vilar, 2018; Hu et al., 2019) with entity constraints to guide the claim generation, by taking the concatenation of all sentence rationales  $R_s$  in  $E/E'$  as input.

Specifically, regarding the list of causal entity tokens  $dc_i = [t_{i,1}, t_{i,2}, \dots, t_{i,j}]$  within each piece

of evidence as disjunctive constraints, where  $t_{i,*} \in T/T'$  denotes the causal entity in the  $i$ -th evidence, and  $j$  is the number of the entities, we acquire the conjunction constraint of the beam search by combining of all disjunctive constraints,

$$CONS = dc_1 \wedge dc_2 \wedge \dots \wedge dc_n, \quad (1)$$

where  $n$  is the number of evidence. The conjunctive constraint during decoding encourages the generated claim  $c$  to contain at least one causal entity from each piece of evidence, thus ensuring factual consistency with the multi-hop evidence. After repeated generation, we generate  $k$  ( $k = 10$  in our experiments) candidate counterfactual claims  $C' = \{c'_1, c'_2, \dots, c'_k\}$  for each instance.

**Post-Checking** The claim generation model can be noisy, potentially leading to the non-reversed predictions of a claim  $c'$  given  $E$ . To ascertain the label flipping between claim  $c'$  and  $c$ , i.e.,  $(c'_i|_{i=1}^k, E) \rightarrow y' \neq y$ , by taking the concatenation of each candidate counterfactual claim  $c'_i$  with its corresponding original evidence  $E$  as input, we use the same three-way fact verification model as in Ad-Checking to filter the candidate counterfactual claims. We retain those candidate claims in  $C'$  that yield a predicted label  $y' \neq y$ .

**Discussion** Claim generation can also be done by very large language models (LLMs) (e.g., ChatGPT (OpenAI, 2022)) with in-context learning (Brown et al., 2020; Wei et al., 2022). However, since our editing may introduce inconsistencies with common sense, we empirically find that the edited evidence  $E'$  is more likely to conflict with the internal knowledge of LLMs, thus leading to the irrelevant content or even failure in generating the claim  $c'$ . Thus, we choose the fine-tuned generation models.

### 3.4 Filtering

Unlike prior work that relies on a curated set of minimal edits (e.g., Yang et al. (2021)), the strategy in our *Generator* maybe over-generate claim  $c'$  with over diverse semantic shift compared to  $c$ . Thus, following Paranjape et al. (2022), we use post-hoc filtering with two modules on generated claims  $C'$  to ensure the minimal semantic (Keane and Smyth, 2020) and topic perturbation compared to the original claim  $c$ .

**Semantic Filtering** The MoverScore (Zhao et al., 2019), which combines the contextualized representations with the Earth Mover distance (Rubner

et al., 2000), measures the semantic similarity between two sentences. We thus use this metric to calculate *semantic fidelity score* between each counterfactual claim in  $C'$  and its corresponding original claim  $c$ , evaluating the semantic change between these two claims.

**Entity Filtering** We introduce the *entity fidelity score* by calculating the overlap rate of entities between strings of claim ( $c', c$ ) pair. This allows us to ensure topic consistency between  $c'$  and  $c$ , filtering out the irrelevant claims from a topic perspective (Si et al., 2021).

One generated claim  $c' \in C'$  with the highest sum score over *semantic fidelity score* and *entity fidelity score* is retained for each instance. Finally, our RACE produces the counterfactual data for each instance  $(c, E, y)$  in the dataset, including  $(c', E, y')$  and  $(c, E', y')$ .

## 4 Experiments

**Datasets** We generate counterfactual data for HOVER<sup>2</sup> training set (Jiang et al., 2020), a multi-hop dataset with facts sourced from Wikipedia. We evaluate the model generalization on three types of development sets, (I) In-domain setting (sourced from Wikipedia), including FEVER (Thorne et al., 2018) and FEVEROUS (Aly et al., 2021). (II) Out-of-domain setting (sourced from specific domains), including PolitiHop (political news) (Ostrowski et al., 2021), SCIFACT (scientific articles) (Wadden et al., 2020), HealthVer (Sarrouti et al., 2021) and PubHealth (public health) (Kotonya and Toni, 2020). (III) Challenge setting (contrastive data), including FM2 (Eisenschlos et al., 2021) and VITAMINC (Schuster et al., 2021). Details and statistics of datasets are presented in Appendix C.

**Baselines** We use three types of baselines to augment the HOVER training set, (I) Data augmentation method: EDA (Wei and Zou, 2019). (II) Counterfactual data augmentation methods: CrossAug (Lee et al., 2021) and POLYJUICE (Wu et al., 2021). (III) LLMs: GPT-3 (text-davinci-003) (Brown et al., 2020) and ChatGPT (gpt-3.5-turbo-0301) (OpenAI, 2022). More details are presented in Appendix D.

<sup>2</sup>Since the HOVER dataset contains explicit multi-hop correlation among evidence based on different reasoning type, we choose it to generate counterfactuals and report results in this paper.

Source of data	$ D_{train} $	In-domain			Out-of-domain				Challenge	
		HOVER	FEVER	FEVEROUS	PolitiHop	SCIFACT	HealthVer	PubHealth	FM2	VITAMINC
None	18,171	82.55	76.70	69.43	48.74	62.77	54.98	53.01	61.51	67.05
EDA	36,342	82.55	73.60	68.22	<b>54.62</b>	62.77	53.68	45.99	60.56	59.63
CrossAug	29,174	82.28	65.92	70.06	<b>54.62</b>	57.98	49.24	39.15	56.12	61.66
POLYJUICE	25,190	81.10	76.43	67.94	45.38	57.98	54.65	46.28	57.14	62.29
GPT-3	24,171	80.75	72.30	67.56	51.26	64.36	49.46	42.91	61.25	62.21
ChatGPT	24,171	80.13	77.77	69.04	44.54	60.64	51.84	45.00	50.74	<b>68.14</b>
our RACE (BART)	24,398	82.78	76.07	70.63	47.06	61.17	46.97	42.81	59.17	59.54
our RACE (GPT-2)	23,645	82.53	77.15	66.07	45.38	<b>65.43</b>	54.87	53.52	<b>62.36</b>	67.88
our RACE (T5-large)	26,638	<b>83.18</b>	<b>78.11</b>	<b>71.55</b>	47.06	62.77	<b>55.84</b>	<b>56.59</b>	61.16	67.71
our RACE (T5-base)	26,917	83.15	75.05	70.50	52.94	<b>65.43</b>	55.41	53.52	62.19	66.50
-CONS	28,468	82.53	73.93	70.09	48.74	59.04	52.71	49.06	62.28	67.31
-EDIT	28,359	80.75	71.50	68.13	54.62	60.64	52.92	47.47	61.33	62.72
-EDIT&CONS	27,682	83.00	76.84	70.69	43.70	60.11	52.60	53.42	60.05	64.74
$w(c, E', REF)$										
our RACE (BART)	27,909	<b>83.33</b>	76.65	69.16	41.18	59.57	54.00	44.20	61.42	66.16
our RACE (GPT-2)	27,156	82.78	75.31	70.52	51.26	62.77	51.62	51.83	59.97	62.18
our RACE (T5-large)	30,149	82.90	<b>78.69</b>	69.29	47.90	64.89	55.41	52.03	61.08	66.31
our RACE (T5-base)	30,428	82.63	76.73	70.90	<b>57.14</b>	60.11	55.63	47.87	61.33	67.66

Table 2: Fact verification accuracy of various data augmentation methods on different development sets in three settings.  $|D_{train}|$  shows the total number of training instances, including 18,171 original HOVER training instances.  $w(c, E', REF)$  denotes the incorporation of counterfactual instances ( $c, E', REF$ ) into the training set. -CONS denotes the use of beam search instead of constrained beam search in claim generation. -EDIT denotes that the evidence editing stage is skipped, and counterfactual claims are generated directly from the original evidence for each original instance. The best of the main results are marked in bold. The results with further improvement in model performance after the incorporation of ( $c, E', REF$ ) are boxed.

**Implementation Details** In the experiments, we fine-tune a basic multi-hop fact verification model, an additional RoBERTa-base (Liu et al., 2019), on the original training data ( $c, E, y$ ) and the counterfactual data generated by each method. The model is evaluated on the development set of different datasets.

For the basic multi-hop fact verification model, we concatenate the claim and all evidence as input sequence, and limit its maximum length to 130. We set the batch size to 4 and optimize the model through a cross entropy loss using the AdamW optimizer (Loshchilov and Hutter, 2019) with the learning rate of  $1e-5$ . For claim generation, we conduct experiments with four generation models: BART-base (Lewis et al., 2020), T5-base, T5-large (Raffel et al., 2020) and GPT-2 (Radford et al., 2019). The beam size is 30 and the max length of generated text is 96.

## 5 Results and Discussion

### 5.1 Main Results

Neglecting the logical relationships within the correlated input results in a failure to generate counterfactual evidence  $E'$  for baselines. Thus, we mainly compare the effects of the counterfactual

data ( $c', E, y'$ ) generated by the different methods. Meanwhile, we also report the results after incorporating ( $c, E', REF$ ) into the training set (bottom of Table 2).

**Out-of-domain Setting** Table 2 shows the effects of the data generated by RACE and baselines on the OOD generalization. We can observe that, (I) RACE significantly improves model performance on PolitiHop, SCIFACT and PubHealth compared to the results without data augmentation, and outperforms baselines on almost all OOD datasets, demonstrating the effectiveness of our augmentation strategy for multi-hop fact verification task. (II) RACE significantly outperforms POLYJUICE, showing that the general-purpose CDA method, designed for tasks without requiring complex reasoning on the input, fails to achieve acceptable results on multi-hop fact verification task, and even impairs the OOD generalization. (III) The counterfactual data generated by LLMs provides little improvement in OOD generalization, demonstrating that CDA for multi-hop fact verification task remains challenging for LLMs by using the in-context learning alone. (IV) The incorporation of ( $c, E', REF$ ) further improves the model generalization to a certain extent on PolitiHop, indicating

Original Instance	
<b>Claim</b>	The 1994 British romantic comedy that Charlotte Ninon Coleman played Scarlett in featured the song "Love".
<b>Evidence</b>	1. [Reg Presley] He wrote the song "Love Is All Around", which was featured in the films "Four Weddings and a Funeral" and "Love Actually". 2. [Charlotte Coleman] Charlotte Ninon Coleman (3 April 1968 – 14 November 2001) was an English actress best known for playing Scarlett in the film "Four Weddings and a Funera", Jess in the television drama "Oranges Are Not the Only Fruit", and her childhood roles of Sue in "Worzel Gummidge" and the character Marmalade Atkins. 3. [Four Weddings and a Funeral] Four Weddings and a Funeral is a 1994 British romantic comedy film directed by Mike Newell.
<b>Label</b>	SUPPORTS
Counterfactual Claims	
<b>CrossAug</b>	The 1994 British romantic comedy that Charlotte Ninon Coleman played Scarlett in featured the song " <b>Love</b> ".
<b>POLYJUICE</b>	The 1994 British romantic comedy that <b>did not win</b> Charlotte Ninon Coleman played Scarlett in featured the song "Love".
<b>ChatGPT</b>	The 1994 <b>American</b> romantic comedy that Charlotte Ninon Coleman played Scarlett in featured the song "Love".
<b>our RACE (T5-base)</b>	<b>Marmalade Atkins directed</b> the 1948 British romantic comedy that <b>Reg Presley</b> played <b>Charlotte Coleman</b> in. <b>It</b> featured the song "Love".

Table 3: Examples of counterfactual claims on HOVER training set derived by different methods. The difference between the counterfactual claim and the original claim is highlighted in blue. See Table 5 in Appendix A for the corresponding edited evidence and more examples.

Method	Flip. $\uparrow$	Flu. $\downarrow$	Sim. $\uparrow$	Div. $\uparrow$	M.h. $\uparrow$
<b>CrossAug</b>	0.3138	209.34	<b>0.6100</b>	2.24	<b>0.6090</b>
<b>POLYJUICE</b>	0.6066	195.18	0.5969	1.20	0.5960
<b>GPT-3</b>	0.3970	96.84	0.5873	1.47	0.5865
<b>ChatGPT</b>	0.4160	107.94	0.5906	1.73	0.5898
<b>RACE (T5-large)</b>	0.9402	<b>55.03</b>	0.5770	11.22	0.5763
<b>RACE (T5-base)</b>	<b>0.9457</b>	55.81	0.5770	11.19	0.5763
<b>-FILTER</b>	0.8388	58.83	0.5773	<b>13.01</b>	0.5766

Table 4: Automatic intrinsic evaluation results. For **Flip rate (Flip.)**, we use a RoBERTa-based classifier fine-tuned on the HOVER training set to calculate the verification accuracy of the instance  $(c', E, y')$ . For **Fluency (Flu.)**, following previous work (Atanasova et al., 2020; He et al., 2023), we use the perplexity scores calculated by GPT-2 to evaluate the fluency of  $c'$ . For **Similarity (Sim.)**, we calculate the MoverScore between  $c'$  and  $c$ . For **Diversity (Div.)**, following Rani et al. (2023), we use the inverse of the BLEU score (Papineni et al., 2002) to measure dissimilarity between  $c'$  and  $c$ . For **Multi hop (M.h.)**, we employ MoverScore to calculate the average semantic similarity between  $c'$  and  $e_i$ , where  $e_i \in E$ , to evaluate the coherence like He et al. (2023). **-FILTER** denotes the evaluation of all the generated claims before post-checking and filtering stage. The best results are marked in bold.

that the edited evidence still remains multi-hop correlated and reasonable.

**Challenge Setting** Comparing the results on challenging contrastive datasets, as Table 2 shows, training with RACE data improves the fact verification accuracy, while almost all the baselines degrade the performance of the model. This phenomenon confirms that our method improves model robustness to spurious correlations. Additionally, the incorporation of the  $(c, E', REF)$  yields no improvement in verification accuracy, probably because these datasets are constructed in response to

the elimination of spurious correlations between features in **claim** and labels.

**In-domain Setting** As shown in Table 2, RACE improves the model performance on in-domain data, while most baselines tend to degrade it. Notably, our method has the most significant improvement on the FEVEROUS development set, which requires four pieces of true evidence to verify each claim on average. This further demonstrates the effectiveness of our method for multi-hop fact verification task.

## 5.2 Ablation Study

We conduct ablation studies on evidence editing and claim generation stage to verify the reasonableness of causal entities in token rationales. All the experiments are conducted based on RACE (T5-base).

Firstly, we use ordinary beam search instead of constrained beam search during the claim generation stage (i.e., **-CONS** in Table 2). The results in Table 2 reveal that a significant performance decrease occurs on both in-domain and OOD data. It might be explained by constraints based on entities in token rationales, which allow the generated claim to be multi-hop and topic consistent with the original claim, resulting in a more efficient counterfactual. In contrast, we note a slight improvement on the challenge datasets, which might be attributed to the shorter length of claims in both datasets (each claim contains about 13 words on average).

Then, we skip the evidence editing stage and directly generate the counterfactual claims for all the instances (i.e., **-EDIT** in Table 2) by a T5-base language model. The model is fine-tuned on FEVEROUS to generate claims that are supported or re-

futed by the input evidence via setting the prefix. As shown in Table 2, the accuracy decreases substantially on almost all datasets, except for PolitiHop. It can be explained by the fact that political news typically focuses on event information rather than entity information, hence entity-based evidence editing fails to improve model generalization on PolitiHop.

Finally, we further remove both the constrained beam search and evidence editing stage (i.e., -*CONS&EDIT* in Table 2). A significant decrease in accuracy is observed on both OOD and challenge data, which demonstrates that the proposed evidence editing based on rationales and claim generation based on entities are crucial for improving the generalization and robustness of the multi-hop fact verification models.

### 5.3 Intrinsic Evaluation

For further analysis of the quality of the generated counterfactual claims, following Chemmen-gath et al. (2022) and Dixit et al. (2022), we automatically and manually evaluate the generated counterfactual claims according to the following five criteria: (I) *Flip rate* (*Flip.*), measuring if the label of the generated claim is flipped based on the original evidence; (II) *Fluency* (*Flu.*), measuring whether the generated claim is grammatically correct and semantically meaningful; (III) *Diversity* (*Div.*), reflecting the linguistic diversity of the generated claim compared to the original claim; (IV) *Similarity* (*Sim.*), measuring the degree of semantic similarity between the generated claim and the original claim, where we use MoverScore (Zhao et al., 2019) instead of Levenshtein edit distance (Levenshtein et al., 1966) in the automatic evaluation to balance with diversity; (V) *Multi hop* (*M.h.*), measuring whether the generated claim is multi-hop and relevant to the evidence.

**Automatic Evaluation** For a fair comparison, the claims generated before and after the post-checking and filtering are compared with the baselines separately. As shown in Table 4, RACE outperforms baselines significantly in terms of flip rate, diversity, and fluency. It demonstrates the ability of RACE to generate fluent and *linguistically diverse* counterfactual claims based on the edited evidence, while keeping *label flipping* and *logical relationships* with the original evidence. Moreover, the counterfactual claim after the filtering stage achieves a higher flip rate and fluency score com-

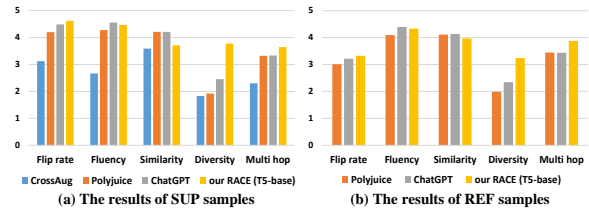


Figure 2: The results of human evaluation, where 1 indicates a complete breach of the criteria and 5 indicates full compliance. The inter-rate agreement measured by Krippendorff’s  $\alpha$  (Krippendorff, 2011) is 0.54.

pared to the one before filtering, which illustrates the necessity of the filtering stage for generating high-quality counterfactual data. For automatic evaluation of *Multi hop*, we follow He et al. (2023) to use MoverScore to evaluate the multi hop of counterfactuals. And all methods achieve comparable results. However, we argue that this is compromised since its solely semantic comparison cannot reflect whether all the evidence can be aggregated as a whole to verify the counterfactual claim.

**Manual Evaluation** To address the limitations of the automatic evaluation, we adopt the human evaluation to qualify the counterfactuals from different aspects. Specifically, we randomly select 30 *SUP* instances and 30 *REF* instances and ask three postgraduate students with an NLP background to score counterfactual claims in a likert scale of 1 to 5 according to the above criteria. Since CrossAug can only generate counterfactuals for *SUP* instances, we compare the results on *SUP* and *REF* instances separately. The evaluation results are shown in Figure 2. It can be observed that RACE well outperforms baselines, particularly in terms of diversity, which illustrates the ability of RACE to generate *human-readable*, *diverse*, and *label-flipping* counterfactual claims. Meanwhile, entity constraint-based generation enables RACE to generate multi-hop claims.

Overall, both the automatic and manual evaluation results show the effectiveness of RACE from different aspects for multi-hop fact verification task.

### 5.4 Qualitative Evaluation

Table 3 presents an example of the original instance and the counterfactual claims generated by different methods. The words that differ from the original claim are highlighted. It can be observed that RACE generates a linguistically diverse and flu-



ent counterfactual claim, and the original label is successfully flipped. Obviously, the counterfactual claim generated by RACE can be combined with the original evidence to form a valid multi-hop fact verification instance, which is logical and can be verified according to the given evidence. Moreover, the claim generated by RACE is semantically and lexically similar to the original claim, benefiting causal entities in multi-hop rationales. Nevertheless, the baselines tend to simply modify the original claim, despite the use of language models. As shown in Table 3, most of the baselines (including LLMs), prefer to add “not” to the original claim or make antonym substitutions. Such modifications make the counterfactual claims lexically similar to the original claim, but are not valid for multi-hop fact verification and cannot generate a diverse and logical counterfactual claim (as evidenced by lower flip rate and diversity in Table 4 and Figure 2).

### 5.5 Effect of Generation Models

We adopt different generation models to test the effect of the generation ability on our method, which aims to illustrate the independence of our proposed method from a particular generation model (i.e., Generation Model-Agnostic). As shown in Table 2, compared to the baselines, our RACE yields a comparable or improved performance based on different generation models, especially the results based on T5-base and T5-large. Besides, We empirically find that different generation models have more prominent performance on specific datasets, e.g., GPT-2 on SCIFACT and FM2 datasets, and T5 on 6 datasets.

To explore the effect of the number of parameters, we further compare the results based on T5-base and T5-large. As Table 4 and 2 shows, compared to T5-base, counterfactuals generated by fine-tuned T5-large are more fluent and linguistically diverse, and further improve the model performance on most datasets. This illustrates that it is possible to further improve the effectiveness of our method by using a more powerful generation model. Thus, for the choice of the generation model, we recommend choosing the powerful possible generation model in the absence of the priors to the data.

## 6 Conclusion

We present a novel rationale-sensitive pipeline counterfactual data augmentation method (RACE) to generate *logical*, *diverse*, and *label-flipping*

counterfactuals for multi-hop fact verification task. An Explain-Edit-Generate architecture is constructed to generate diverse and logical counterfactual claims based on the rationales. Then, a filter process with two modules is employed to further regularize semantic and topic consistency. Experimental results reveal the improvement in OOD generalization and robustness of the proposed method. Intrinsic evaluation and qualitative evaluation of counterfactual claims show that RACE can generate linguistically diverse and label-flipping counterfactual data while preserving logical relationships.

### Limitations

As multi-hop fact verification is a relatively complex reasoning task, designing an effective method to generate counterfactuals for this task requires a consideration of the logical relationships between the claim and the evidence and between multiple pieces of evidence, making our proposed method more complex and cumbersome. Meanwhile, the use of heuristic rules in the editing process results in the inability to generalize to other tasks and the need to recreate the rules. In addition, the prompts given to LLMs for generating counterfactual claims can be further elaborated, e.g., using chain-of-thought, to exploit more potential of LLMs on CDA for multi-hop fact verification task.

In the future, due to the flexible generation of LLMs, we will explore the construction of effective prompts to generate counterfactuals for multi-hop fact verification using the Chain-of-Thought.

### Acknowledgement

The authors would like to thank the anonymous reviewers for their insightful comments. This work is funded by the National Natural Science Foundation of China (62176053) and supported by the Big Data Computing Center of Southeast University. YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1, EP/V020579/2).

### References

- Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. [Where is your evidence: Improving fact-checking by justification modeling](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.
- Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopou-

- Ios, Oana Cocarascu, and Arpit Mittal. 2021. [Feverous: Fact extraction and verification over unstructured and structured information](#). In *Proceedings of the NeurIPS Track on Datasets and Benchmarks*.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. [Guided open vocabulary image captioning with constrained beam search](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 936–945.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. [Fact checking with insufficient evidence](#). *Transactions of the Association for Computational Linguistics*, 10:746–763.
- Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. [Generating label cohesive and well-formed adversarial claims](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3168–3177.
- Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. [Brenda: Browser extension for fake news detection](#). In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2117–2120.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Saneem Chemmengath, Amar Prakash Azad, Ronny Luss, and Amit Dhurandhar. 2022. [Let the CAT out of the bag: Contrastive attributed explanations for text](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7190–7206.
- Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyong Zhou, and William Yang Wang. 2020. [Tabfact: A large-scale dataset for table-based fact verification](#). In *International Conference on Learning Representations*.
- Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2):103219.
- Michiel de Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, and William W. Cohen. 2021. [Mention memory: incorporating textual knowledge into transformers through entity mention attention](#). *CoRR*, abs/2110.06176.
- Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [CORE: A retrieve-then-edit framework for counterfactual data generation](#). In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 2964–2984.
- Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. [Fool me twice: Entailment from Wikipedia gamification](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365.
- Jacob Eisenstein. 2022. [Informativeness and invariance: Two perspectives on spurious correlations in natural language](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4326–4331.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 1307–1323.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that require simple lexical inferences](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 650–655.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.
- Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. [Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation](#). In *Proceedings of the ACM Web Conference*, pages 2698–2709.
- Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. [NeuroCounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation](#). In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 5056–5072.
- J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. [Improved lexically constrained decoding for translation and monolingual rewriting](#).

- In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–850.
- 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 *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 3441–3460.
- Nitish Joshi and He He. 2022. **An investigation of the (in)effectiveness of counterfactually augmented data**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3668–3681.
- Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. **Learning the difference that makes a difference with counterfactually-augmented data**. In *International Conference on Learning Representations*.
- Mark T Keane and Barry Smyth. 2020. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *Proceedings of International Conference on Case-Based Reasoning Research and Development*, pages 163–178.
- Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. **More bang for your buck: Natural perturbation for robust question answering**. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 163–170.
- Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. **Baleen: Robust multi-hop reasoning at scale via condensed retrieval**. In *Advances in Neural Information Processing Systems*, pages 27670–27682.
- Neema Kotonya and Francesca Toni. 2020. **Explainable automated fact-checking for public health claims**. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7740–7754.
- Klaus Krippendorff. 2011. **Computing krippendorff’s alpha-reliability**.
- Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. **Crossaug: A contrastive data augmentation method for debiasing fact verification models**. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pages 3181–3185.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, pages 707–710.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. **Roberta: A robustly optimized BERT pretraining approach**. *CoRR*, abs/1907.11692.
- Ilya Loshchilov and Frank Hutter. 2019. **Decoupled weight decay regularization**. In *International Conference on Learning Representations*.
- OpenAI. 2022. **Introducing chatgpt**.
- Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. **Multi-hop fact checking of political claims**. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3892–3898.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. **Retrieval-guided counterfactual generation for QA**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1670–1686.
- Matt Post and David Vilar. 2018. **Fast lexically constrained decoding with dynamic beam allocation for neural machine translation**. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1314–1324.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. **Stanza: A python natural language processing toolkit for many human languages**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1:9.
- 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. *The Journal of Machine Learning Research*, 21:5485–5551.
- Tathagata Raha, Mukund Choudhary, Abhinav Menon, Harshit Gupta, KV Srivatsa, Manish Gupta, and Vasudeva Varma. 2023. Neural models for factual inconsistency classification with explanations. *arXiv preprint arXiv:2306.08872*.

- Anku Rani, SM Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Factify-5wqa: 5w aspect-based fact verification through question answering. *arXiv preprint arXiv:2305.04329*.
- Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover’s distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99.
- Mourad Sarrouiti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. **Evidence-based fact-checking of health-related claims**. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. **Get your vitamin C! robust fact verification with contrastive evidence**. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643.
- 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 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 3419–3425.
- Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. **Topic-aware evidence reasoning and stance-aware aggregation for fact verification**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pages 1612–1622.
- Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023a. Consistent multi-granular rationale extraction for explainable multi-hop fact verification. *arXiv preprint arXiv:2305.09400*.
- Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023b. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 2023. **Multi2Claim: Generating scientific claims from multi-choice questions for scientific fact-checking**. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 2652–2664.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. **FEVER: a large-scale dataset for fact extraction and VERification**. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. **The spread of true and false news online**. *Science*, 359(6380):1146–1151.
- 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 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550.
- David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. **MultiVerS: Improving scientific claim verification with weak supervision and full-document context**. In *Proceedings of Findings of the Association for Computational Linguistics: NAACL*, pages 61–76.
- Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14024–14031.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. **Chain of thought prompting elicits reasoning in large language models**. In *Advances in Neural Information Processing Systems*.
- Jason Wei and Kai Zou. 2019. **EDA: Easy data augmentation techniques for boosting performance on text classification tasks**. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 6382–6388.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. **Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pages 6707–6723.
- Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. **Discovering invariant rationales for graph neural networks**. In *International Conference on Learning Representations*.
- Linyi Yang, Jiazheng Li, Pdraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. **Exploring the efficacy of automatically generated counterfactuals for sentiment analysis**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pages 306–316.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. **MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance**. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 563–578.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. **Generalizing to the future: Mitigating entity bias in fake news detection**. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125.

## A Evidence Editing

Table 5 shows examples of the evidence edited by RACE. We can observe that rationale- and entity-based editing enables the edited evidence to still retain multi-hop correlation with each other and present a completely different fact from the original evidence. Hence, the claim generator can generate logical, fluent, and linguistically diverse counterfactual claims based on the edited evidence.

## B Checking Module

For the ad- and post-checking module, we fine-tune a RoBERTa-base classifier to filter invalid edited evidence and counterfactual claims, respectively. To improve the quality of the retained data, we fine-tune it on the *SUP*, *REF*, and *NEI* instances rather than just the *SUP* and *REF* instances.

Considering that we perform CDA on HOVER training set during the experiment while no *NEI* instances are available in HOVER, we first conduct data augmentation on HOVER dataset to incorporate *NEI* instances by perturbing existing instances. Specifically, for a random instance in HOVER, we randomly remove one piece of true evidence or randomly pair the claim with the evidence of another instance. To avoid imbalance classes, we randomly select half of the *SUP* instances and half of the *REF* instances for perturbation and each perturbation strategy is employed with equal probability. Finally, the fine-tuned RoBERTa-base classifier has 81.23% on label accuracy of claim verification on *NEI* augmented HOVER development set. The statistics of *NEI* augmented HOVER are shown in Table 6.

Other implementation details are the same as the fact verification model in the OOD generalization experiment described in Section 4.

## C Datasets

- **HOVER** (Jiang et al., 2020), a dataset for multi-hop fact verification, which challenges models to extract relevant evidence from several Wikipedia articles and verify whether the claim is SUPPORTED or REFUTED by the evidence. We construct the dataset following Khattab et al. (2021), where each claim is associated with five pieces of evidence.
- **FEVER** (Thorne et al., 2018), a large-scale fact verification dataset with the claims generated by altering sentences extracted from Wikipedia. The claims in FEVER are classified as SUPPORTS, REFUTES or NOT ENOUGH INFO (NEI) by annotators and more than 87% of them only require information from a single Wikipedia article. We remove the instances with NEI label and only retain the other two classes of instances in our experiments.
- **FEVEROUS** (Aly et al., 2021), a large-scale multi-hop fact verification dataset consisting of claims verified against Wikipedia pages and labeled as SUPPORTS, REFUTES or NOT ENOUGH INFO (NEI). Each claim has evidence in the form of sentences and/or cells from tables on Wikipedia. Following Chen et al. (2020) and Si et al. (2023b), we employ the simple table linearization template to generate contextualized sequence representations for table evidence. We remove the instances with NEI label and only retain the other two classes of instances in our experiments.
- **PolitiHop** (Ostrowski et al., 2021), a multi-hop fact verification dataset of real-world claims with manual annotations of evidence from PolitiFact articles. The labels include FALSE, HALF-TRUE and TRUE. In our experiments, we remove the instances with HALF-TRUE label and only retain the other two classes of instances.
- **SCIFACT** (Wadden et al., 2020), a scientific fact verification dataset of 1.4K expert-written scientific claims paired with evidence. As with the above dataset, we only retain the instances with SUPPORTS and REFUTES labels to evaluate the model.

Original Instance	
<b>Claim</b>	The 1994 British romantic comedy that Charlotte Ninon Coleman played Scarlett in featured the song “Love”.
<b>Evidence</b>	<ol style="list-style-type: none"> <li>[Reg Presley] He wrote the song “Love Is All Around”, which was featured in the films “Four Weddings and a Funeral” and “Love Actually”.</li> <li>[Charlotte Coleman] Charlotte Ninon Coleman (3 April 1968 – 14 November 2001) was an English actress best known for playing Scarlett in the film “Four Weddings and a Funera”, Jess in the television drama “Oranges Are Not the Only Fruit”, and her childhood roles of Sue in “Worzel Gummidge” and the character Marmalade Atkins.</li> <li>[Four Weddings and a Funeral] Four Weddings and a Funeral is a 1994 British romantic comedy film directed by Mike Newell.</li> </ol>
<b>Label</b>	SUPPORTS
Edited Evidence	
<b>Edired Evidence</b>	<ol style="list-style-type: none"> <li>[Mike Newell] He wrote the song “Love Is All Around”, which was featured in the films “Four Weddings and a Funeral” and “Love Actually”.</li> <li>[Reg Presley] <a href="#">Reg Presley (3 August 1987 – 26 June 2000)</a> was an English actress best known for playing <a href="#">Charlotte Coleman</a> in the film “Four Weddings and a Funera”, <a href="#">Scarlett</a> in the television drama “Oranges Are Not the Only Fruit”, and her childhood roles of Sue in “Worzel Gummidge” and the character <a href="#">Jess</a>.</li> <li>[Four Weddings and a Funeral] Four Weddings and a Funeral is a <a href="#">1948</a> British romantic comedy film directed by <a href="#">Marmalade Atkins</a>.</li> </ol>
Counterfactual Claims	
<b>CrossAug</b>	The 1994 British romantic comedy that Charlotte Ninon Coleman played Scarlett in featured the song “ <a href="#">Love</a> ”.
<b>POLYJUICE</b>	The 1994 British romantic comedy that <a href="#">did not win</a> Charlotte Ninon Coleman played Scarlett in featured the song “Love”.
<b>ChatGPT</b>	The 1994 <a href="#">American</a> romantic comedy that Charlotte Ninon Coleman played Scarlett in featured the song “Love”.
<b>our RACE</b>	<a href="#">Marmalade Atkins</a> directed the 1948 British romantic comedy that <a href="#">Reg Presley</a> played <a href="#">Charlotte Coleman</a> in. It featured the song “Love”.
Original Instance	
<b>Claim</b>	Bruce Geller who died in 1978 developed American television detective show Mannix.
<b>Evidence</b>	<ol style="list-style-type: none"> <li>[Mannix] Created by Richard Levinson and William Link and developed by executive producer Bruce Geller, the title character, Joe Mannix, is a private investigator.</li> <li>[Bruce Geller] Bruce Bernard Geller (October 13, 1930 – May 21, 1978) was an American lyricist, screenwriter, director, and television producer.</li> </ol>
<b>Label</b>	SUPPORTS
Edited Evidence	
<b>Edired Evidence</b>	<ol style="list-style-type: none"> <li>[Mannix] Created by <a href="#">Joe Mannix</a> and Richard Levinson and developed by executive producer <a href="#">William Link</a>, the title character, <a href="#">Bruce Geller</a>, is a private investigator.</li> <li>[<a href="#">William Link</a>] <a href="#">William Link (December 14, 1898 – April 30, 1977)</a> was an American lyricist, screenwriter, director, and television producer.</li> </ol>
Counterfactual Claims	
<b>CrossAug</b>	Bruce Geller who <a href="#">passed away</a> in 1978 developed American television detective show Mannix.
<b>POLYJUICE</b>	Bruce Geller who died in 1978, <a href="#">did not</a> developed American television detective show Mannix.
<b>ChatGPT</b>	Bruce Geller, who <a href="#">passed away</a> in <a href="#">1985</a> , developed the American television detective show Mannix.
<b>our RACE</b>	<a href="#">The executive producer of American television detective show Mannix</a> died in <a href="#">1877</a> . <a href="#">The show was created by Joe Mannix and Richard Levinson</a> .

Table 5: Examples of edited evidence and counterfactual claims on HOVER training set. The differences from the original instance are highlighted in blue.

Augmented HOVER	Num.SUP	Num.REF	Num.NEI	Total
Train	11,023	7,148	9,086	27,572
Dev	2,000	2,000	2,000	6,000

Table 6: The statistics of augmented HOVER with NEI instances. Num.SUP, Num.REF and Num.NEI are the number of SUP instances, REF instances, and NEI instances, respectively.

- **HealthVer** (Sarrouti et al., 2021), an evidence-based fact verification dataset for health-related claims, where the relations between each piece of evidence and the associated claim are manually annotated as SUPPORT,

REFUTE, and NEUTRAL. We remove the instances with the NEUTRAL label. As the evidence provided by HealthVer contains several sentences, we split it into multiple pieces of evidence to simulate a multi-hop scenario.

- **PubHealth** (Kotonya and Toni, 2020), a 4-way classification dataset for explainable fact verification with gold standard explanations by journalists in the public health setting. We only retain the instances with TRUE and FALSE labels, and the explanation provided is split into separate sentences as multiple pieces of evidence.

Given an original claim with corresponding evidence and label (SUPPORTS or REFUTES), generate a counterfactual claim based on the evidence, taking care to ensure that the generated counterfactual claim is as **similar** as possible to the original claim, while being aware of linguistic **diversity** and the **change** of labels.

**Example:**

**Claim:** Bettany Hughes, an English historian scholar, born May 15th, 1967, presented "The Spartans".

**Evidence:**

The Spartans (documentary): "The Spartans" was a 3-part historical documentary series first broadcast on UK terrestrial Channel 4 in 2003, presented by Bettany Hughes.

Bettany Hughes: Bettany Hughes ( born May 15 , 1967 ) is an English historian, author, and broadcaster.

**Label:** SUPPORTS

**Generate a counterfactual claim:**

"The Spartans" is a documentary presented by Bettany Hughes, an American historian scholar born on March 24, 1980.

**Claim:** The writer Norman Alfred William Lindsay enjoyed boxing, but the author of The Hundred Secret Senses did not.

**Evidence:**

Amy Tan: Amy Tan ( born February 19, 1952 ) is an American writer whose works explore mother-daughter relationships and the Chinese American experience.

The Hundred Secret Senses: The Hundred Secret Senses is a bestselling 1995 novel by Chinese-American writer Amy Tan.

Norman Lindsay: Norman Alfred William Lindsay ( 22 February 1879 – 21 November 1969 ) was an Australian artist, etcher, sculptor, writer, editorial cartoonist, scale modeller, and an accomplished amateur boxer.

**Label:** SUPPORTS

**Generate a counterfactual claim:**

Table 7: An example of prompt given to GPT-3 and ChatGPT for generating counterfactual claims.

Dataset	Num.SUP	Num.REF	Total
HOVER Dev	2,000	2,000	4,000
FEVER Dev	6,666	6,666	13,332
FEVEROUS Dev	3,908	3,481	7,389
PolitiHop Dev	21	98	119
SCIFACT Dev	124	64	188
HealthVer Dev	533	391	924
PubHealth Dev	628	544	1,172
FM2 Dev	596	573	1,169
VITAMINC Dev	31,484	22,528	54,012

Table 8: The statistics of the datasets we used in our experiments.

- **FM2** (Eisenschlos et al., 2021), a large-scale dataset of challenging claim-evidence pairs collected through a fun multi-player game which encourages adversarial instances and drastically lowers the number of the instances with "shortcuts". All the claims need to be verified  $\in \{ \text{SUPPORTS}, \text{REFUTES} \}$ .
- **VITAMINC** (Schuster et al., 2021), a large-scale contrastive fact verification dataset, where each contrastive claim is manually written by annotators based on Wikipedia revisions. We only retain the instances with SUP-

PORTS and REFUTES labels in our experiments.

We only test the performance of the basic multi-hop fact verification model on the development set of the above datasets in our experiments. The statistics are shown in Table 8.

## D Baselines

In our experiments, we compare our method with the following baselines.

- **EDA** (Wei and Zou, 2019), a data augmentation method that applies four simple operations, including synonym replacement, random insertion, random swap, and random deletion, to original sentences to generate new instances.
- **CrossAug** (Lee et al., 2021), a counterfactual data augmentation method that employs a two-stage augmentation pipeline to generate contrastive claims and evidence from existing *SUP* instances.
- **POLYJUICE** (Wu et al., 2021), a general-purpose counterfactual generator based on fine-tuned GPT-2 that allows for control over perturbation types and locations.

- **GPT-3** (text-davinci-003) (Brown et al., 2020), a large autoregressive language model with superb few-shot and in-context learning capabilities.
- **ChatGPT** (gpt-3.5-turbo-0301) (OpenAI, 2022), a powerful GPT-3 based model which is trained to follow an instruction in a prompt and provide a detailed response.

For EDA<sup>3</sup> and CrossAug<sup>4</sup>, all the experimental setups of them are followed from the original papers and all hyperparameters are set to the same values as in the official code. For POLYJUICE<sup>5</sup>, we set the control code to “negation”, the beam size to 10, and generate one counterfactual claim for each original claim. All the inputs to the above baselines are only the original claim.

For GPT-3 and ChatGPT, we make use of the APIs provided by OpenAI<sup>6</sup> for generating counterfactual claims and design a prompt with a task introduction and demonstration as input, as shown in the Table 7.

---

<sup>3</sup>[https://github.com/jasonwei20/eda\\_nlp](https://github.com/jasonwei20/eda_nlp)

<sup>4</sup><https://github.com/minwhoo/CrossAug>

<sup>5</sup><https://github.com/tongshuangwu/polyjuice>

<sup>6</sup><https://openai.com/product>