Zero-shot Faithful Factual Error Correction

Kung-Hsiang Huang[♠] **Hou Pong Chan**[♡] **Heng Ji**[♠]
[♠]Department of Computer Science, University of Illinois Urbana-Champaign
[♡]Faculty of Science and Technology, University of Macau
[♠]{khhuang3, hengji}@illinois.edu
[♡]hpchan@um.edu.mo

Abstract

Faithfully correcting factual errors is critical for maintaining the integrity of textual knowledge bases and preventing hallucinations in generative models. Drawing on humans' ability to identify and correct factual errors, we present a zero-shot framework that formulates questions about input claims, looks for correct answers in the given evidence, and assesses the faithfulness of each correction based on its consistency with the evidence. Our zero-shot framework outperforms fully-supervised approaches, as demonstrated by experiments on the FEVER and SciFact datasets, where our outputs are shown to be more faithful. More importantly, the decomposability nature of our framework inherently provides interpretability. Additionally, to reveal the most suitable metrics for evaluating factual error corrections, we analyze the correlation between commonly used metrics with human judgments in terms of three different dimensions regarding intelligibility and faithfulness.[1]

1 Introduction

The task of correcting factual errors is in high demand and requires a significant amount of human effort. The English Wikipedia serves as a notable case in point. It is continually updated by over 120K editors, with an average of around six factual edits made per minute[2]. Using machines to correct factual errors could allow the articles to be updated with the most current information automatically. This process, due to its high speed, can help retain the integrity of the content and prevent the spread of false or misleading information.

In addition, the hallucination issues have been shown to be a prime concern for neural models,

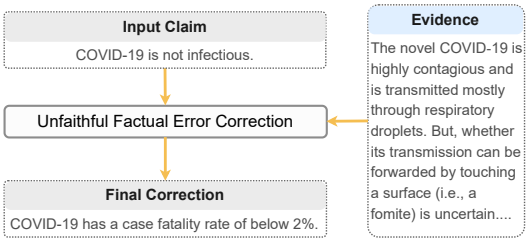

Figure 1: An example of a factual but unfaithful correction leading to misleading information. While it is technically true that the majority of people infected with COVID-19 will recover, there is no information in the evidence that supports the final correction. Additionally, when this statement is taken out of context, it could mislead people to believe that COVID-19 is not dangerous and that there is no need for precautions, which is false. A factual and faithful correction is "COVID-19 is highly contagious.".

where they are prone to generate content factually inconsistent with the input sources due to the unfaithful training samples (Maynez et al., 2020) and the implicit "knowledge" it learned during pre-training (Niven and Kao, 2019). Factual error correction can be used in both pre-processing and post-processing steps to rectify the factual inconsistencies in training data and generated texts, respectively. This can help build trust and confidence in the reliability of language models.

Prior work typically formulates factual error correction as a sequence-to-sequence task, either in a fully supervised or in a distantly supervised manner (Shah et al., 2020; Thorne and Vlachos, 2021). While these approaches have made great strides in generating fluent and grammatically valid corrections, they only focus on the aspect of *factuality*: *whether the outputs are aligned with facts*. Little emphasis was placed on *faithfulness*: *the factual consistency of the outputs with the evidence*. Faithfulness is critical in this task as it indicates whether a generated correction reflects the information we intend to update. If faithfulness is not ensured,

[1]The code and data have been made publicly available: https://github.com/khuangaf/ZeroFEC
[2]https://en.wikipedia.org/wiki/Wikipedia:Statistics

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5660–5676

this could lead to the spread of misleading content, causing serious consequences. Figure 1 shows a concrete example. In the context of automatically updating textual knowledge bases, the topic of an unfaithful output would likely deviate much from that of the expected correction. Therefore, such an edit is not desirable, even if it is factual.

In this work, we present the first study on the *faithfulness* aspect of factual error correction. To address faithfulness, we propose a *zero-shot* factual error correction framework (ZEROFEC), inspired by how humans verify and correct factual errors. When humans find a piece of information suspicious, they tend to first identify potentially false information units, such as noun phrases, then ask questions about each information unit, and finally look for the correct answers in trustworthy evidence (Saeed et al., 2022; Chen et al., 2022). Following a similar procedure, ZEROFEC breaks the factual error correction task into five sub-tasks: (1) *claim answer generation*: extracting all information units, such as noun phrases and verb phrases, from the input claim; (2) *question generation*: generating question given each *claim answer* and the original claim such that each *claim answer* is the answer to each generated question; (3) *question answering*: answering each generated question using the evidence as context; (4) *QA-to-claim*: converting each pair of generated question and answer to a declarative statement; (5) *correction scoring*: evaluating corrections based on their faithfulness to the evidence, where faithfulness is approximated by the entailment score between the evidence and each candidate correction. The highest-scoring correction is selected as the final output. An overview of our framework is shown in Figure 2. Our method ensures the corrected information units are derived from the evidence, which helps improve the faithfulness of the generated corrections. In addition, our approach is *naturally interpretable* since the questions and answers generated directly reflect which information units are being compared with the evidence.

Our contributions can be summarized as follows:

- We propose ZEROFEC, a factual error correction framework that effectively addresses faithfulness by asking questions about the input claim, seeking answers in the evidence, and scoring the outputs by faithfulness.
- Our approach outperforms all prior methods, including fully-supervised approaches trained on 58K instances, in ensuring faithfulness on two factual error correction datasets, FEVER (Thorne et al., 2018) and SCIFACT (Wadden et al., 2020).
- We analyze the correlation of human judgments with automatic metrics to provide intuition for future research on evaluating the faithfulness, factuality, and intelligibility of factual error corrections.

2 Task

In Thorne and Vlachos (2021)'s setting, retrieved evidence is used, which means the model may be able to correct factual errors, even though there is no supporting information in the evidence. In this case, although the prediction is considered correct, the model is hallucinating, which is not a desired property. Additionally, due to the way data was collected, they require systems to alter the input claim even if the input claim is already faithful to the evidence. We argue that *no edit is needed for claims that are faithful to the evidence*.

To address these shortcomings, our setup aims to edit a claim using a given piece of grounded evidence that supports or refutes the original claim (see Figure 2). Using gold-standard evidence avoids the issue where a system outputs the correct answer by chance due to hallucinations. In our setting, a system must be faithful to the evidence to correct factual errors, allowing us to evaluate system performance more fairly. Furthermore, we require the model not to edit the original claim if it is already factually consistent with the provided evidence.

Concretely, the input to our task is a claim $\mathcal{C}$ and a piece of gold-standard evidence $\mathcal{E}$ that supports or refutes $\mathcal{C}$. The goal of factual error correction is to produce a corrected claim $\hat{\mathcal{C}}$ that fixes factual errors in $\mathcal{C}$ while being faithful to $\mathcal{E}$. If $\mathcal{C}$ is already supported by $\mathcal{E}$, models should output the original claim (i.e. $\hat{\mathcal{C}} = \mathcal{C}$).

3 Proposed Methods

Our framework, ZEROFEC, faithfully corrects factual errors using question-answering and entailment. Specifically, we represent the input claim $\mathcal{C}$ as question-answer pairs $\{(Q_1, A_1^{\mathcal{C}}), ..., (Q_n, A_n^{\mathcal{C}})\}$ such that each question Q_i reflects the corresponding information unit $A_i^{\mathcal{C}}$, such as noun phrases and adjectives (§3.1 and §3.2). Based on each question Q_i, we look for an

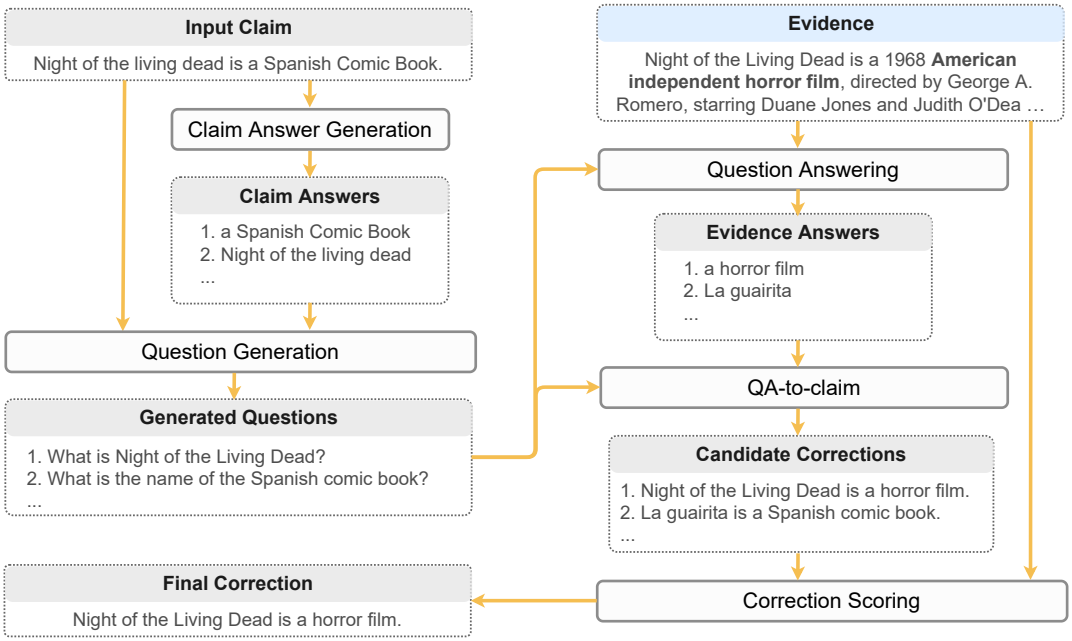

Figure 2: An overview of our framework. First, given an input claim, we generate the *claim answers* by enumerating all information units in the input claim. Second, conditioned on each extracted answer and the input claim, a question is generated. Third, each question is then fed to a question answering model to produce an *evidence answer* using the given evidence as context. Fourth, using a sequence-to-sequence approach, each *evidence answer* and the corresponding question are transformed into a statement, which serves as a *candidate correction*. Finally, the *final correction* is produced by scoring candidate corrections based on faithfulness.

answer $A_i^{\mathcal{E}}$ in the given evidence $\mathcal{E}$ using a learned QA model (§3.3). Each candidate correction S_i is obtained by converting the corresponding pair of Q_i and $A_i^{\mathcal{E}}$ into a declarative statement (§3.4). This guarantees that the corrected information units we replace factual errors with are derived from the evidence and thus ensures high faithfulness. The final output of ZEROFEC is the S_i with the highest faithfulness score computed by an entailment model (§3.5). An overview of our framework is shown in Figure 2.

One major challenge that makes our task more difficult than prior studies on faithfulness (Wang et al., 2020; Fabbri et al., 2022a) is that we need to handle more diverse factual errors, such as negation errors and errors that can only be abstractively corrected. For instance, in the second example of in Table 2, the QA model should output "Yes" as the answer, which cannot be produced by extractive QA systems. To address this issue, we adopt abstractive QG and QA models that can handle diverse question types and train our QA-to-claim model on multiple datasets to cover cases that cannot be handled by extractive systems. The following subsections illustrate the details of each component in our framework.

3.1 Claim Answer Generation

The goal of claim answer generation is to identify information units in the input claim that may be unfaithful to $\mathcal{E}$. We aim to maximize the recall in this step since the missed candidates cannot be recovered in later steps. Therefore, we extract all noun chunks and named entities using Spacy[3] and extract nouns, verbs, adjectives, adverbs, noun phrases, verb phrases using Stanza[4]. Additionally, we also extract negation terms, such as "not" and "never", from the input claim. We name the extracted information units *claim answers*, denoted as $A^{\mathcal{C}} = \{A_1^{\mathcal{C}}, A_2^{\mathcal{C}}, ..., A_n^{\mathcal{C}}\}$.

3.2 Question Generation

Upon *claim answers* are produced, we generate questions that will be later used to look for correct information units in the evidence. Questions are generated conditioned on the *claim answers* using the input claim as context. We denote the question generator as $\mathcal{G}$. Each *claim answer* $A_i^{\mathcal{C}}$ is

[3] https://spacy.io/
[4] https://stanfordnlp.github.io/stanza/

concatenated with the input claim $\mathcal{C}$ to generate a question $Q_i = \mathcal{G}(A_i^{\mathcal{C}}, \mathcal{C})$. We utilize MixQG (Murakhovs'ka et al., 2022) as our question generator $\mathcal{G}$ to cover the wide diversity of factual errors and candidates extracted. MixQG was trained on nine question generation datasets with various answer types, including boolean, multiple-choice, extractive, and abstractive answers.

3.3 Question Answering

The question answering step identifies the correct information units $A_i^{\mathcal{E}}$ corresponding to each question Q_i in the given evidence $\mathcal{E}$. Our QA module answers questions from the question generation steps with the given evidence as context. Let $\mathcal{F}$ denote our QA model. We feed the concatenation of a generated question and the evidence to the QA model to produce an *evidence answer* $A_i^{\mathcal{E}} = \mathcal{F}(Q_i, \mathcal{E})$. UnifiedQA-v2 (Khashabi et al., 2022) is used as our question answering model. UnifiedQA-v2 is a T5-based (Raffel et al., 2020b) abstractive QA model trained on twenty QA datasets that can handle diverse question types.

3.4 QA-to-Claim

After questions and answers are generated, we transform each pair of question and answer into a declarative statement, which serves as a candidate correction that will be scored in the next step. Previous studies on converting QAs to claims focus on extractive answer types only (Pan et al., 2021). To accommodate diverse types of questions and answers, we train a sequence-to-sequence model that generates a claim given a question-answer pair on three datasets: QA2D (Demszky et al., 2018) for extractive answers, BoolQ (Clark et al., 2019) for boolean answers, and SciTail (Khot et al., 2018) for covering scientific domain QAs. Note that samples in BoolQ do not contain converted declarative statements. Using Stanza's constituency parser, we apply heuristics to transform all QAs to their declarative forms in BoolQ. Our QA-to-claim model is a T5-base fine-tuned on these three datasets. Concretely, let $\mathcal{M}$ denote our QA-to-claim model. $\mathcal{M}$ takes in a *generated question* $\mathcal{Q}_i$ and an *evidence answer* $A_i^{\mathcal{E}}$ as inputs and outputs a statement $S_i = \mathcal{M}(Q_i, A_i^{\mathcal{E}})$.

3.5 Correction Scoring

The final correction is produced by scoring the faithfulness of each candidate correction from the previous steps w.r.t. the evidence. We use entailment score to approximate faithfulness. Here, DocNLI (Yin et al., 2021) is used to compute such document-sentence entailment relations. DocNLI is more generalizable than other document-sentence entailment models, such as FactCC (Kryscinski et al., 2020), since it was trained on five datasets of various tasks and domains. Conventional NLI models trained on sentence-level NLI datasets, such as MNLI (Williams et al., 2018), are not applicable since previous work has found that these models are ill-suited for measuring entailment beyond the sentence level (Falke et al., 2019). In addition, to prevent the final correction from deviating too much from the original claim, we also consider ROUGE-1 scores, motivated by Wan and Bansal (2022). The final metric used for scoring is the sum of ROUGE-1 score[5] and DocNLI entailment score. Formally,

$$\mathcal{V}(S_i) = \text{DocNLI}(S_i, \mathcal{E}) + \text{ROUGE-1}(S_i, \mathcal{C}) \quad (1)$$

$$\mathcal{C}' = \underset{S_i}{\text{argmax}}\, \mathcal{V}(S_i), \quad (2)$$

where $\mathcal{C}'$ is the final correction produced by our framework. Furthermore, to handle cases where the input claim is already faithful to the evidence, we include the input claim in the candidate correction list to be scored.

3.6 Domain Adaptation

During the early stage of our experiments, we found that our proposed framework did not perform well in correcting factual errors in biomedical claims. This results from the fact that our QA and entailment models were not fine-tuned on datasets in the biomedical domain. To address this issue, we adapt UnifiedQA-v2 and DocNLI on two biomedical QA datasets, PubMedQA (Jin et al., 2019) and BioASQ (Tsatsaronis et al., 2015), by further fine-tuning them for a few thousand steps. We later show that this simple domain adaptation technique successfully improves our overall factual error correction performance on a biomedical dataset without decreasing performance in the Wikipedia domain (see §5.1).

4 Experimental Setup

4.1 Datasets

We conduct experiments on two English datasets, FEVER and SciFact. FEVER (Thorne and Vla-

[5] https://pypi.org/project/py-rouge/

chos, 2021) is repurposed from the corresponding fact-checking dataset (Thorne et al., 2018) that consists of evidence collected from Wikipedia and claims written by humans that are supported or refuted by the evidence. Similarly, SCIFACT is another fact-checking dataset in the biomedical domain (Wadden et al., 2020). We repurpose it for the factual error correction task using the following steps. First, we form faithful claims by taking all claims supported by evidence. Then, unfaithful claims are generated by applying Knowledge Base Informed Negations (Wright et al., 2022), a semantic altering transformation technique guided by knowledge base, to a subset of the faithful claims. Appendix A shows detailed statistics.

4.2 Evaluation Metrics

Our evaluation focuses on faithfulness. Therefore, we adopt some recently developed metrics that have been shown to correlate well with human judgments in terms of faithfulness. BARTScore (Yuan et al., 2021) computes the semantic overlap between the input claim and the evidence by calculating the logarithmic probability of generating the evidence conditioned on the claim. FactCC (Kryscinski et al., 2020) is an entailment-based metric that predicts the faithfulness probability of a claim w.r.t. the evidence. We report the average of the CORRECT probability across all samples. In addition, we consider QAFACTEVAL (Fabbri et al., 2022a), a recently released QA-based metric that achieves the highest performance on the SUMMAC factual consistency evaluation benchmark (Laban et al., 2022). Furthermore, we also report performance on SARI (Xu et al., 2016), a lexical-based metric that has been widely used in the factual error correction task (Thorne and Vlachos, 2021; Shah et al., 2020).

4.3 Baselines

We compare our framework with the following baseline systems. **T5-FULL** (Thorne and Vlachos, 2021) is a fully-supervised model based on T5-base (Raffel et al., 2020a) that generates the correction conditioned on the input claim and the given evidence. **MASKCORRECT** (Shah et al., 2020) and **T5-DISTANT** (Thorne and Vlachos, 2021) are both distantly-supervised methods that are composed of a masker and a sequence-to-sequence (seq2seq) corrector. The masker learns to mask out information units that are possibly false based on a learned fact verifier or an explanation model (Ribeiro et al., 2016) and the seq2seq corrector learns to fill in the masks with factual information. The biggest difference is in the choice of seq2seq corrector. **T5-DISTANT** uses T5-base, while **MASKCORRECT** utilizes a two-encoder pointer generator. For zero-shot baselines, we selected two post-hoc editing frameworks that are trained to remove hallucinations from summaries, **REVISEREF** (Adams et al., 2022) and **COMPEDIT** (Fabbri et al., 2022b). **REVISEREF** is trained on synthetic data where hallucinating samples are created by entity swaps. **COMPEDIT** learns to remove factual errors with sentence compression, where training data are generated with a separate perturber that inserts entities into faithful sentences.

4.4 Implementation Details

No training is needed for ZEROFEC. As for ZEROFEC-DA, we fine-tune UNIFIEDQA-V2 and DOCNLI on the BIOASQ and PUBMEDQA datasets for a maximum of 5,000 steps using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 3e-6 and a weight decay of 1e-6. During inference time, all generative components use beam search with a beam width of 4.

5 Results

5.1 Main Results

Table 1 summarizes the main results on the FEVER and SCIFACT datasets. Both ZEROFEC and ZEROFEC-DA achieve significantly better performance than the distantly-supervised and zero-shot baselines. More impressively, they surpass the performance of the fully-supervised model on most metrics, even though the fully-supervised model is trained on 58K samples in the FEVER experiment. The improvements demonstrate the effectiveness of our approach in producing faithful factual error correction by combining question answering and entailment predictions. In addition, even though our domain adaptation technique is simple, it successfully boosts the performance on the SCIFACT dataset while pertaining great performance on the FEVER dataset. The first example in Table 2 illustrates an instance where domain adaptation fixes an error made by ZEROFEC. The absence of domain adaptation results in incorrect predictions by ZEROFEC, as DocNLI assigns a significantly lower entailment score to the correct candidate "Clathrin stabilizes the spindle fiber apparatus during mitosis phase." and a higher score to the wrong candidate "Clathrin stabilizes the spindle apparatus during

Method	FEVER				SciFact			
	SARI (%)	BS	QFE	FC (%)	SARI (%)	BS	QFE	FC (%)
Fully-supervised								
T5-Full	35.50	-2.74	1.40	41.91	35.07	-3.12	1.23	50.17
Distantly-supervised								
MaskCorrect	25.66	-4.48	0.67	20.12	15.21	-4.31	0.54	34.92
T5-Distant	36.01	-2.90	1.12	32.28	20.08	-3.51	0.99	44.77
Zero-shot								
ReviseRef	20.52	-5.27	0.30	26.00	17.53	-4.58	0.97	**52.44**
CompEdit	25.51	-2.83	1.23	39.46	25.41	-3.31	1.12	50.62
ZeroFEC (Ours)	39.16*	**-2.58***	**2.06***	**47.08***	29.67	-3.22	1.12	47.84
ZeroFEC-DA (Ours)	**40.65***	-2.67*	2.03*	45.75*	**31.93**	**-3.21**	**1.30***	50.10

Table 1: Main results on the FEVER and SciFact datasets. BS denotes BARTScore, QFE denotes QAFactEval, and FC denotes FactCC. ZeroFEC-DA is our framework with the QA and entailment components further fine-tuned on biomedical QA datasets. Among distantly-supervised and zero-shot results, the best scores per metric are marked in **boldface**. Models achieving performance better than the fully-supervised model are marked in gray. Statistical significance over previous best methods computed with the paired bootstrap procedure (Berg-Kirkpatrick et al., 2012) are indicated with * ($p < .01$).

anaphase?", indicating poor entailment assessment. With domain adaptation, ZeroFEC-DA resolves this issue by enabling DocNLI to approximate faithfulness more accurately.

It is true that ZeroFEC-DA requires additional training, which is different from typical zero-shot methods. However, the key point remains that our framework does not require any task-specific training data. Hence, our approach still offers the benefits of zero-shot learning by not requiring any additional training data beyond what was already available for the question answering task, a field with much richer resources compared to the fact-checking field.

5.2 Qualitative Analysis

To provide intuition for our framework's ability to produce faithful factual error corrections, we manually examined 50 correct and 50 incorrect outputs made by ZeroFEC on the FEVER dataset. The interpretability of ZeroFEC allows for insightful examinations of the outputs. Among the correct samples, our framework produces faithful corrections because all intermediate outputs are accurately produced rather than "being correct by chance". For the incorrect outputs, we analyze the source of mistakes, as shown in Figure 3. The vast majority of failed cases result from DocNLI's failure to score candidate corrections faithfully. In addition to the mediocre performance of DocNLI, one primary reason is that erroneous outputs from other components would not be considered mistakes so long as the correction scoring module determines the resulting candidate corrections unfaithful to the evidence. A possible solution to improve DocNLI is to further fine-tune it on synthetic data generated by perturbing samples in FEVER and SciFact. Examples of correct and incorrect outputs are presented in Table 7 and Table 8 of Appendix D, respectively.

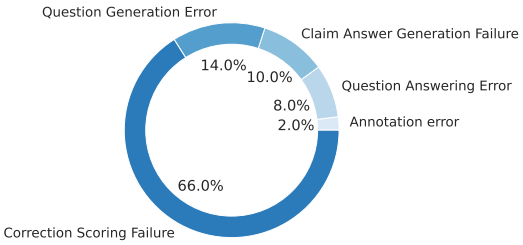

Figure 3: Distributions of errors.

5.3 Human Evaluation

To further validate the effectiveness of our proposed method, we recruited three graduate students who are not authors to conduct human evaluations on 100 and 40 claims from FEVER and SciFact, respectively. For each claim, human judges are presented with the ground-truth correction, the gold-standard evidence, and output produced by a factual error correction system and tasked to assess the quality of the correction with respect to three dimensions. ***Intelligibility*** evaluates the fluency of the correction. An intelligible output is free of grammatical mistakes, and its meaning must be

Example 1	
Input claim: Clathrin stabilizes the spindle fiber apparatus during anaphase. **Evidence**: ...but is shut down during mitosis, when clathrin concentrates at the spindle apparatus... **Gold correction**: Clathrin stabilizes the spindle fiber apparatus during mitosis.	
Claim answer: anaphase **Evidence answer**: mitosis **DocNLI + ROUGE-1**: 0.0165 + 0.8235	**Generated question**: Clathrin stabilizes the spindle fiber apparatus during what phase? **Candidate correction**: Clathrin stabilizes the spindle fiber apparatus during mitosis phase. **ZEROFEC's output**: Clathrin stabilizes the spindle apparatus during anaphase?
Claim answer: anaphase **Evidence answer**: mitosis **DocNLI + ROUGE-1**: 0.9999 + 0.8235	**Generated question**: Clathrin stabilizes the spindle fiber apparatus during what phase? **Candidate correction**: Clathrin stabilizes the spindle fiber apparatus during mitosis phase. **ZEROFEC-DA's output**: Clathrin stabilizes the spindle fiber apparatus during mitosis phase.
Example 2	
Input claim: Fuller House (TV series) won't air on Netflix. **Evidence**: Fuller House is an American family sitcom and sequel to the 1987-95 television series Fuller House, airing as a Netflix original series... **Gold correction**: Fuller House (TV series) airs on Netflix.	
Claim answer: won't air on Netflix **Evidence answer**: Yes **DocNLI + ROUGE-1**: 0.7222 + 0.7143	**Generated question**: Does Fuller House air on Netflix? **Candidate correction**: Fuller House airs on Netflix. **ZEROFEC's output**: Fuller House airs on Netflix.
T5-DISTANT's output: Fuller House (TV series) isn't airing on HBO.	

Table 2: Example outputs from different approaches. The outputs from our framework are directly interpretable, as the generated questions and answers reflect which information units in the input claim are erroneous and which information in the evidence supports the final correction. We only show the generated answers and questions directly related to the gold correction. In the first example, ZEROFEC-DA corrects a mistake made by ZEROFEC thanks to domain adaptation. In the second example, ZEROFEC successfully produces a faithful factual error correction, whereas the output of T5-DISTANT, the distantly-supervised baseline, is factual yet unfaithful to the evidence.

understandable by humans without further explanation. ***Factuality*** considers whether the input claim is aligned with facts. Systems' output can be factual and semantically different from the gold correction as long as it is consistent with the world's knowledge. ***Faithfulness*** examines whether the input is factually consistent with the given evidence. Note that a faithful output must be factual since we assume all evidence is free of factual error. To evaluate the annotation quality, we compute the inter-annotator agreement. Krippendorff's Alpha (Krippendorff, 2011) is 68.85%, which indicates a moderate level of agreement. Details of our human evaluation can be found in Appendix B.

The human evaluation results are demonstrated in Table 3. We observe that: (1) ZEROFEC and ZEROFEC-DA achieve the best overall performance in *Factuality* and *Faithfulness* on both datasets, even when compared to the fully-supervised method, suggesting that our approach is the best in ensuring faithfulness for factual error correction. (2) Our domain adaptation for the biomedical domain surprisingly improves faithfulness and factuality in the Wikipedia domain (i.e. FEVER). This suggests that fine-tuning the components of our framework on more datasets helps improve robustness in terms of faithfulness. (3) Factual output produced by ZEROFEC and ZEROFEC-DA are always faithful to the evidence, preventing the potential spread of misleading information caused by factual but unfaithful corrections. The second example in Table 2 demonstrates an instance of factual but unfaithful correction made by baseline models. Here, the output of T5-DISTANT is unfaithful since the evidence does not mention whether Fuller House airs on HBO. In fact, although Fuller House was not on HBO when it premiered, it was later accessible on HBO Max. Therefore, the correction produced by T5-DISTANT is misleading.

Method	FEVER			SCIFACT		
	Intel.	*Fact.*	*Faith.*	*Intel.*	*Fact.*	*Faith.*
T5-FULL	0.983	0.516	0.509	0.972	0.683	0.610
T5-DISTANT	0.891	0.471	0.412	0.628	0.186	0.116
ZEROFEC	**0.951**	0.797	0.797	0.826	0.413	0.413
ZEROFEC-DA	0.893	**0.835**	**0.835**	**0.953**	**0.628**	**0.628**

Table 3: Human evaluation on the FEVER and SCIFACT datasets. *Intel.* denotes *Intelligibility*, *Fact.* denotes *Factuality*, and *Faith.* denotes *Faithfulness*.

5.4 Correlation with Human Judgments

Recent efforts on faithfulness metrics have been mostly focusing on the summarization task. No prior work has studied the transferability of these metrics to the factual error correction task. We seek to bridge this gap by showing the correlation between the automatic metrics used in Table 1 and the human evaluation results discussed in §5.3. Using Kendall's Tau (Kendall, 1938) as the correlation

Metric	FEVER			SCIFACT		
	Intel.	*Fact.*	*Faith.*	*Intel.*	*Fact.*	*Faith.*
SARI	0.017	**0.370**	**0.383**	-0.026	**0.379**	**0.412**
BARTSCORE	**0.137**	0.071	0.104	**0.104**	0.118	0.119
QAFACTEVAL	-0.045	0.360	0.379	0.084	0.234	0.272
FACTCC	0.053	0.203	0.225	-0.119	-0.073	-0.076

Table 4: Correlation between automatic metrics and human judgments on the FEVER and SCIFACT datasets computed using Kendall's Tau.

measure, the results are summarized in Table 4.

We have the following observations. (1) SARI is the most consistent and reliable metric for evaluating *Factuality* and *Faithfulness* across two datasets. Although the other three metrics developed more recently demonstrate high correlations with human judgments of faithfulness in multiple summarization datasets, their transferability to the factual error correction task is limited due to their incompatible design for this particular task. For example, QA-based metrics like QAFACTEVAL are less reliable for evaluating faithfulness in this task due to their inability to extract a sufficient number of answers from a single-sentence input claim. In contrast, summaries in summarization datasets generally consist of multiple sentences, enabling the extraction of a greater number of answers. To validate this, we analyzed the intermediate outputs of QAFACTEVAL. Our analysis confirms that it extracts an average of only 1.95 answers on the FEVER dataset, significantly lower than the more than 10 answers typically extracted for summaries. (2) Across the two datasets, the correlations between all automatic metrics and *Intelligibility* are low. The extremely high proportion of intelligible outputs may explain the low correlation. (3) The correlation for learning-based metrics, including QAFACTEVAL and FACTCC, drop significantly when applied to SCIFACT. This is likely caused by the lack of fine-tuning or pre-training with biomedical data.

6 Related Work

6.1 Factual Error Correction

An increasing number of work began to explore factual error correction in recent years, following the rise of fact-checking (Thorne et al., 2018; Wadden et al., 2020; Gupta and Srikumar, 2021; Huang et al., 2022b) and fake news detection (Shu et al., 2020; Fung et al., 2021; Wu et al., 2022; Huang et al., 2022a). Shah et al. (2020) propose a distant supervision learning method based on a masker-corrector architecture, which assumes access to a learned fact verifier. Thorne and Vlachos (2021) created the first factual error correction dataset by repurposing the FEVER (Thorne et al., 2018) dataset, which allows for fully-supervised training of factual error correctors. They also extended Shah et al. (2020)'s method with more advanced pre-trained sequence-to-sequence models. Most recently, Schick et al. (2022) proposed PEER, a collaborative language model that demonstrates superior text editing capabilities due to its multiple text-infilling pre-training objectives, such as planning and realizing edits as well as explaining the intention behind each edit[6].

6.2 Faithfulness

Previous studies addressing faithfulness are mostly in the summarization field and can be roughly divided into two categories, evaluation and enhancement. Within faithfulness evaluation, one line of work developed entailment-based metrics by training document-sentence entailment models on synthetic data (Kryscinski et al., 2020; Yin et al., 2021) or human-annotated data (Ribeiro et al., 2022; Chan et al., 2023), or applying conventional NLI models at the sentence level (Laban et al., 2022). Another line of work evaluates faithfulness by comparing information units extracted from summaries and input sources using QA (Wang et al., 2020; Deutsch et al., 2021). There is a recent study that integrates QA into entailment by feeding QA outputs as features to an entailment model (Fabbri et al., 2022a). We combine QA and entailment by using entailment to score the correction candidates produced by QA.

Within faithfulness enhancement, some work improves factual consistency by incorporating auxiliary losses into the training process (Nan et al., 2021; Cao and Wang, 2021; Tang et al., 2022; Huang et al., 2023). Some other work devises factuality-aware pre-training and fine-tuning objectives to reduce hallucinations (Wan and Bansal, 2022). The most similar to our work are studies that utilize a separate rewriting model to fix hallucinations in summaries. For example, Cao et al. (2020) present a post-hoc corrector trained on synthetic data, where negative samples are created via perturbations. Adams et al. (2022) fix factually inconsistent information in the reference summaries

[6]We are not able to compare with PEER (Schick et al., 2022) as its checkpoints have not been released by the time we ran the experiments.

to prevent the summarization from learning hallucinating examples. Fabbri et al. (2022b) propose a compression-based post-editor to correct extrinsic errors in the generated summaries. By contrast, we leverage the power of QA and entailment together to address faithfulness.

7 Conclusions and Future Work

We have presented ZEROFEC, a zero-shot framework that asks questions about an input claim and seeks answers from the given evidence to correct factual errors faithfully. The experimental results demonstrate the superiority of our approach over prior methods, including fully-supervised methods, as indicated by both automatic metrics and human evaluations. More importantly, the decomposability of ZEROFEC naturally offers interpretability, as the questions and answers generated directly reflect which information units in the input claim are incorrect and why. Furthermore, we reveal the most suitable metric for assessing faithfulness of factual error correction by analyzing the correlation between the reported automatic metrics and human judgments. For future work, we plan to extend our framework to faithfully correct misinformation in social media posts and news articles to inhibit the dissemination of false information. In addition, it may be meaningful to explore extending zero-shot factual error correction to multimedia task settings, such as identifying inconsistencies between chart and text (Zhou et al., 2023).

8 Limitations

Although our approach has demonstrated advantages in producing faithful factual error corrections, we recognize that our approach is not capable of correcting all errors, particularly those that require domain-specific knowledge, as illustrated in Table 3. Therefore, it is important to exercise caution when applying this framework in user-facing settings. For instance, end users should be made aware that not all factual errors may be corrected.

In addition, our approach assumes evidence is given. Although this assumption is also true for applying our method to summarization tasks since the source document is treated as evidence, it does not hold for automatic textual knowledge base updates. When updating these knowledge bases, it is often required to retrieve relevant evidence from external sources. Hence, a reliable retrieval system is required when applying our method to this task.

9 Ethical Considerations

While no fine-tuning is needed for ZEROFEC, its inference time and memory usage are three to four times more than similar-sized baseline systems due to its multi-component architecture, implying higher environmental costs during test time. In addition, the underlying components of our method are based on language models pre-trained on data collected from the internet. These language models have been shown to exhibit potential issues, such as political or gender biases. While we did not observe such biases during our experiments, users of these models should be aware of these issues when applying them.

Acknowledgement

This research is based upon work supported by U.S. DARPA SemaFor Program No. HR001120C0123, DARPA AIDA Program No. FA8750-18-2-0014, and DARPA MIPs Program No. HR00112290105. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Hou Pong Chan was supported in part by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ) and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST).

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6251–6258, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hou Pong Chan, Qi Zeng, and Heng Ji. 2023. Interpretable automatic fine-grained inconsistency detection in text summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022a. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R Fabbri, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu, and Caiming Xiong. 2022b. Improving factual consistency in summarization with compression-based post-editing. *arXiv preprint arXiv:2211.06196*.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022a. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*.

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. 2023. SWING: Balancing coverage and faithfulness for dialogue summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 512–525, Dubrovnik, Croatia. Association for Computational Linguistics.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022b. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. MixQG: Neural question generation with mixed answer types. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1486–1497, Seattle, United States. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Leonardo Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. Factgraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3238–3253. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1736–1746.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*.

Darsh Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8791–8798.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

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

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

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

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

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Mingyang Zhou, Yi Fung, Chris Thomas, Long Chen, Heng Ji, and Shih-fu Chang. 2023. Enhance chart understanding via visual language pre-training on plot table pairs. In *ACL Findings*.

A Dataset Statistics

Details of the dataset statistics are shown in Table 5.

Dataset	# Test Samples	# SUPPORTS	# REFUTES
FEVER	3,882	1,593	2,289
SCIFACT	100	43	57

Table 5: Statistics of FEVER and SCIFACT.

B Human Evaluation Details

In this section, we describe the details of our human evaluation. We recruit three engineering and science graduate students to ensure high-quality evaluation. For each HIT, annotators are provided with an input claim, the corresponding evidence and gold correction, and a predicted correction generated by a model. Based on the presented predictions, annotators are tasked to answer three questions shown on the right segment of the interface, each of which corresponds to *Intelligence*, *Factuality*, and *Faithfulness*. They need to determine whether the predicted correction meets the three criteria according to each prompt. Our human evaluation interface is displayed in Figure 4.

Since the evaluation questions are self-explanatory, we only provide the human evaluators with terminology definitions and multiple examples of how evaluations should be conducted. Terminology is defined as follows:

- **Input claim**: A sentence fed into a factual error correction system.
- **Predicted correction**: The output from the factual error correction system.
- **Gold correction**: Ground-truth label that the system aims to produce.
- **Evidence**: A document that the factual error correction system used to fix factual errors.

We maintain frequent communication with the human evaluators, including answering any questions they may have, to facilitate the evaluation process.

C Ablation Studies

To understand how each component contributes to the performance of ZEROFEC, we conducted ablation studies by replacing a given component in ZEROFEC with other models while keeping all other components the same as ZEROFEC. We report the performance on the FEVER dataset in SARI and QAFACTEVAL since these two metrics demonstrate the highest correlation with human judgments regarding faithfulness. Ablation results are presented in Table 6.

Model/Data Choice	SARI (%)	QAFACTEVAL
Question Generation		
MixQG-base	39.16	**2.06**
T5-base (SQuAD)	**39.19**	2.04
Question Answering		
UnifiedQA-v2-base	**39.16**	2.06
UnifiedQA-base	39.02	**2.09**
T5-base (SQuAD)	30.38	1.02
RoBERTa-base (SQuAD)	31.42	1.11
QA-to-claim		
T5-base (QA2D + BoolQ + SciTail)	**39.16**	**2.06**
T5-base (QA2D)	30.54	1.23
T5-base (SciTail)	29.24	1.19
Correction Scoring		
DocNLI + ROUGE-1	**39.16**	**2.06**
DocNLI	34.56	1.95
FactCC + ROUGE-1	30.54	1.47
FactCC	30.33	1.45

Table 6: Ablation studies on the FEVER dataset. The model used in ZEROFEC is **bolded**.

Effect of Question Generation We compared MixQG with a T5-base model trained on SQuAD (Rajpurkar et al., 2016). The results indicate that the final performance is not significantly affected by the use of either model. Upon further investigation, we surprisingly discovered that despite SQuAD exclusively comprising extractive question answering examples, the T5-base trained on it could generalize to other answer types. For example, given an answer "not" and a claim "Cleopatre is not a queen.", T5-base (SQuAD) generates "Is Cleopatre a queen?". Therefore, the training of MixQG on multiple QA datasets does not yield advantages.

Effect of Question Answering We experimented with an abstractive QA model, UnifiedQA (Khashabi et al., 2020), and two extractive QA models trained on SQuAD. We found that UnifiedQA performs similarly to UnifiedQA-v2, whereas using both extractive QA models leads to significant performance drops. This is likely due to the fact that SQuAD only includes extractive answer types. Although the encoder-decoder architecture of T5-base allows it to output words that do not present in the context, it fails to generate these types of answers. For instance, given a question "Was Cleopatre a

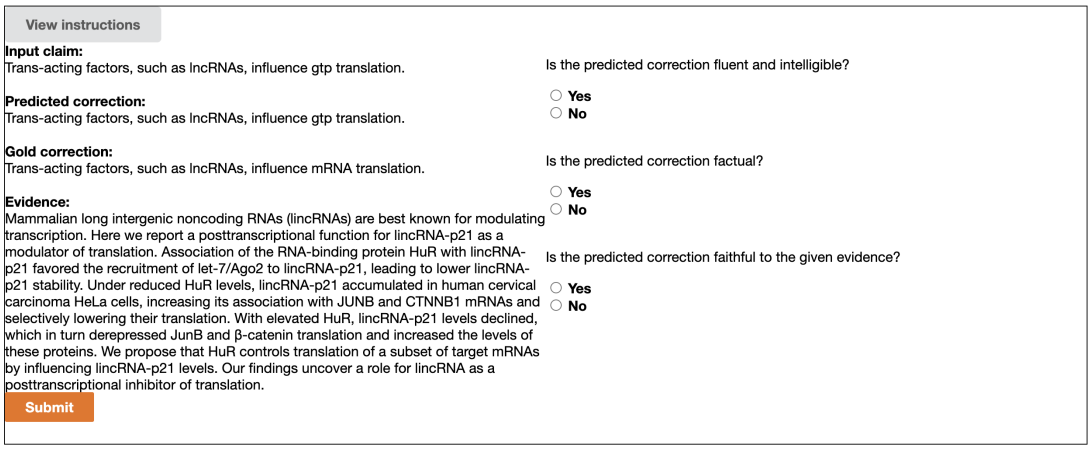

Figure 4: MTurk UI for our human evaluation.

queen." and a context "Cleopatra VII Philopator was Queen of the Ptolemaic Kingdom of Egypt...", T5-base (SQuAD) would output "Queen" instead of "Yes".

Effect of QA-to-claim For QA-to-claim, we ablated different training data while keeping the same model architecture. Similar to our findings in the ablation studies on QA, when T5-base is only trained on QA2D or SciTail, it cannot convert boolean-typed questions and answers to declarative sentences, resulting in a marked decline in performance.

Effect of Correction Scoring We studied other scoring methods, including replacing DocNLI with FactCC and removing ROUGE-1. Using FactCC leads to a great performance drop, suggesting that DocNLI is likely a better approximation of faithfulness than FactCC. Furthermore, incorporating ROUGE-1 into the scoring criteria allows us to select a faithful correction that is most relevant to the input claim. Thus, we observe a huge drop in SARI when ROUGE-1 is removed.

D Additional Qualitative Analysis

As mentioned in §5.1, we analyzed 50 correct and 50 incorrect outputs produced by ZEROFEC. All 50 correct outputs are generated by asking the correct questions, answering correctly using the evidence, and scoring faithfully w.r.t. the evidence. Examples are demonstrated in Table 7. For incorrect outputs, most of the errors are caused by DocNLI's inability to approximate faithfulness, as shown by the last instance in Table 8, even though DocNLI is the state-of-the-art document-sentence entailment model. In addition, annotation errors occur due to how the FEVER dataset was constructed (i.e. for fact-checking purposes). As demonstrated by the first example in Table 8, our correction is faithful to the evidence, and it is also more relevant to the input claim compared to the ground truth. As for errors in the question answering module, most of them are under-specified answers. For example, in the second instance in Table 8, the generated answer "pop music duo" is faithful to the evidence but is under-specified compared to the expected answer "R&B singers".

E Software and Hardware Configurations

All experiments were conducted on a Ubuntu 18.04.6 Linux machine with a single NVIDIA V100. We use PyTorch 1.11.0 with CUDA 10.2 as the Deep Learning framework and utilize Transformers 4.19.2 to load all pre-trained language models.

F Number of Parameters

The number of parameters for each component in ZEROFEC is provided in parentheses: MixQG-base (220M), UnifiedQA-v2-base (220M), QA-to-claim (220M), DocNLI (355M).

G Scientific Artifacts

The licenses for all the models and software used in this paper are listed below in parentheses: Spacy (MIT License), Stanza (Apache License 2.0), MixQG-base (BSD-3-Clause License), UnifiedQA-v2 (Apache License 2.0), T5-base (Apache License 2.0), DocNLI (BSD-3-Clause License), py-ROUGE (Apache License 2.0), FACTCC (BSD-3-Clause License), QAFACTEVAL (BSD-3-Clause License), SARI (GPL-3.0 License), BARTSCORE (Apache License 2.0).

Input claim: University of Chicago Law School is ranked first in the 2016 QS World University Rankings. **Evidence**: The University of Chicago Law School is the graduate school of law at the University of Chicago. It is ranked 12th in the 2016 QA World University Rankings. **Gold correction**: University of Chicago Law School is ranked 12th in the 2016 QS World University Rankings.	
Claim answer: ranked **Evidence answer**: 12th **DocNLI + ROUGE-1**: 0.8867 + 0.9032	**Generated question**: How is the University of Chicago Law School ranked in the 2016 QS World University Rankings? **Candidate correction**: The University of Chicago Law School is 12th ranked in the 2016 QS World University Rankings. **ZEROFEC's output**: The University of Chicago Law School is 12th ranked in the 2016 QS World University Rankings.
Input claim: Simon Pegg was born on February 14th, 1860. **Evidence**: Fuller House is an American family sitcom and sequel to the 1987-95 television series Full House, airing as a Netflix original series ... **Gold correction**: Simon Pegg was born on February 14th, 1970.	
Claim answer: February 14th, 1860 **Evidence answer**: 14 february 1970 **DocNLI + ROUGE-1**: 0.9636 + 0.7500	**Generated question**: When was Simon Pegg born? **Candidate correction**: Simon Pegg was born on 14 february 1970. **ZEROFEC's output**: Simon Pegg was born on 14 february 1970.
Input claim: Caesar is a 1937 adaptation of The Tempest. **Evidence**: Fuller Caesar is the title of Orson Welles's innovative 1937 adaptation of William Shakespeare's Julius Caesar ... **Gold correction**: Caesar is a 1937 adaptation of Julius Caesar.	
Claim answer: Tempest **Evidence answer**: William Shakespeare's Julius Caesar **DocNLI + ROUGE-1**: 0.9649 + 0.6315	**Generated question**: Caesar is a 1937 adaptation of what? **Candidate correction**: Caesar is a 1937 adaptation of William Shakespeare's Julius Caesar. **ZEROFEC's output**: Caesar is a 1937 adaptation of William Shakespeare's Julius Caesar.

Table 7: Example outputs from ZEROFEC on the FEVER dataset where it makes faithful factual error corrections.

Annotation Error	
Input claim: Ragtime was released in 1965. **Evidence**: Ragtime is a 1981 American drama film, ... The film features the final film appearances of James Cagney and Pat O'Brien ... **Gold correction**: Ragtime is the final film appearance of James Cagney.	
Claim answer: released in 1965 **Evidence answer**: 1981 **DocNLI + ROUGE-1**: 0.9636 + 0.7500	**Generated question**: When was Ragtime released? **Candidate correction**: Ragtime was released in 1981. **ZEROFEC's output**: Ragtime was released in 1981.
Under-specified Answer	
Input claim: Sonny & Cher are rock singers. **Evidence**: Sonny & Cher were an American pop music duo ... started their career in the mid-1960s as R&B backing singers ... **Gold correction**: Sonny & Cher are R&B singers.	
Claim answer: rock singers **Evidence answer**: pop music duo **DocNLI + ROUGE-1**: 0.8166 + 0.5000	**Generated question**: Sonny & Cher are what type of singers? **Candidate correction**: Sonny & Cher are a pop music duo. **ZEROFEC's output**: Sonny & Cher are a pop music duo.
Correction Scoring Failure	
Input claim: Johann Wolfgang von Goethe failed to publish Wilhelm meister's Apprenticeship. **Evidence**: ... During this period, Goethe published his second novel, Wilhelm Meister's Apprenticeship ... **Gold correction**: Johann Wolfgang von Goethe published Wilhelm Meister's Apprenticeship	
Candidate correction (A): Johann Wolfgang von Goethe published Wilhelm Meister's Apprenticeship. **Candidate correction (B)**: Johann Wolfgang von Goethe failed to published Wilhelm Meister's Apprenticeship. **ZEROFEC's output**: Johann Wolfgang von Goethe failed to publish Wilhelm meister's Apprenticeship.	**DocNLI + ROUGE-1 (A)**: 0.0203 + 0.9000 **DocNLI + ROUGE-1 (B)**: 0.1011 + 1.0000

Table 8: Example outputs from ZEROFEC on the FEVER dataset where it fails to produce faithful factual error corrections. The three types of errors correspond to mistakes in *annotation*, *question answering*, and *correction scoring*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8.

☑ A2. Did you discuss any potential risks of your work?
Section 9.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1.

☑ A4. Have you used AI writing assistants when working on this paper?
We use Grammarly to check the language/grammar.

B ☑ Did you use or create scientific artifacts?

Appendix G.

☑ B1. Did you cite the creators of artifacts you used?
Section 3 & 4.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix G.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4, Appendix G.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A.

C ☑ Did you run computational experiments?

Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix E & F.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 4 & 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 3 & 4.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5. Appendix B.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B.