Entity-level Factual Adaptiveness of Fine-tuning based Abstractive Summarization Models

Jongyoon Song[1] **Nohil Park**[1] **Bongkyu Hwang**[2] **Jaewoong Yun**[2]
Seongho Joe[2] **Youngjune L. Gwon**[2] **Sungroh Yoon**[1,3,4*]
[1]Data Science & AI Laboratory, Seoul National University, Korea
[2]Samsung SDS, Korea
[3]Deptment of ECE and Interdisciplinary Program in AI, Seoul National University, Korea
[4]ASRI, INMC, and AIIS, Seoul National University, Korea
{coms1580, pnoil2588, sryoon}@snu.ac.kr
{bongkyu.hwang, jw0531.yun, drizzle.cho, gyj.gwon}@samsung.com

Abstract

Abstractive summarization models often generate factually inconsistent content particularly when the parametric knowledge of the model conflicts with the knowledge in the input document. In this paper, we analyze the robustness of fine-tuning based summarization models to the knowledge conflict, which we call *factual adaptiveness*. We utilize pre-trained language models to construct evaluation sets and find that factual adaptiveness is not strongly correlated with factual consistency on original datasets. Furthermore, we introduce a controllable counterfactual data augmentation method where the degree of knowledge conflict within the augmented data can be adjustable. Our experimental results on two pre-trained language models (PEGASUS and BART) and two fine-tuning datasets (XSum and CNN/DailyMail) demonstrate that our method enhances factual adaptiveness while achieving factual consistency on original datasets on par with the contrastive learning baseline.

1 Introduction

Factual consistency is a crucial aspect, especially in abstractive summarization, ensuring that the facts presented in the generated summary align with those in the input document (Maynez et al., 2020; Kryscinski et al., 2020; Huang et al., 2021; Scialom et al., 2021; Fabbri et al., 2022).

Recent summarization models using pre-training and/or fine-tuning of the language model have shown excellent performance in various aspects such as factual consistency (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020; Cao and Wang, 2021; Wan and Bansal, 2022; Roit et al., 2023). There are also studies on large language models (Brown et al., 2020; Ouyang et al., 2022; Chowdhery et al., 2022) for the summarization (Zhang et al., 2023; Adams et al., 2023) or the evaluation of summaries (Luo et al., 2023; Gao et al., 2023).

Previous works have also reported that (large) language models have *parametric knowledge* (Ji et al., 2023; Bang et al., 2023). The parametric knowledge of the language model is known to result in hallucinated contents, particularly when *knowledge conflict* occurs which refers to the mismatch between the knowledge in the document and the parametric knowledge of the model (Longpre et al., 2021; Neeman et al., 2022; Zhou et al., 2023b). Because the hallucination problem degrades factual consistency of summarization models (Maynez et al., 2020; Nan et al., 2021), it is important to study the robustness to knowledge conflict of summarization models.

In abstractive summarization, most previous works have measured the factual consistency using the original document only, which is not sufficient to evaluate the robustness to knowledge conflict (Cao and Wang, 2021; Wan and Bansal, 2022; Wan et al., 2023). There are studies on hallucination problems caused by knowledge conflict in abstractive summarization (Ladhak et al., 2023; Cheang et al., 2023). However, the aforementioned document perturbation strategies do not control the degree of knowledge conflict, which offers valuable insight into the robustness of the summarization models to the knowledge conflict.

In this paper, we define *factual adaptiveness*, the robustness to the knowledge conflict, of fine-tuning based abstractive summarization models. We focus on entity-level knowledge conflict and factual adaptiveness and use counterfactual samples obtained by replacing a single named entity (i.e., original entity) with another named entity (i.e., counterfactual entity).

Unlike previous works on knowledge conflict in question answering, there are two additional considerations in our work (Longpre et al., 2021;

[*] Corresponding author

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 915–929
March 17-22, 2024 ©2024 Association for Computational Linguistics

Neeman et al., 2022). First, we determine which named entity to replace in the reference summary by detecting parametric knowledge. Second, we select the named entity to be replaced with to control knowledge conflict. To address those considerations, we utilize the parametric knowledge of the pre-trained language model (PLM) during the knowledge conflict set construction.

We first analyze the factual adaptiveness of various methods for improving factual consistency on original datasets such as data filtering (Nan et al., 2021), contrastive learning (Cao and Wang, 2021), and advanced decoding (Wan et al., 2023). Our results demonstrate that methods for factual consistency on original datasets do not always effectively mitigate knowledge conflict problems, which indicates that factual consistency on original datasets can be orthogonal to factual adaptiveness.

We next propose a controllable counterfactual data augmentation technique. Specifically, the method constructs counterfactual samples based on a pre-defined degree of knowledge conflict. Experimental results show that our method improves factual adaptiveness effectively and addresses the entity-level hallucination problem caused by knowledge conflict.

Our contributions can be summarized as follows:

- We introduce the factual adaptiveness of fine-tuning based summarization models using a parametric knowledge of a pre-trained language model.
- We demonstrate that factual consistency on original datasets tends to be orthogonal to factual adaptiveness. Specifically, data filtering largely improves factual adaptiveness while advanced decoding and contrastive learning show minimal differences.
- We propose a controllable counterfactual data augmentation method that enhances factual adaptiveness while preserving factual consistency on original datasets.

2 Factual Adaptiveness

In this section, we define and analyze factual adaptiveness of various fine-tuning based summarization models which are known to improve factual consistency. We formulate factual adaptiveness in Section 2.1, and explain the factual adaptiveness evaluation set construction method in Section 2.2. In the remaining text, the term *counterfactual* indicates the presence of knowledge conflict caused by the entity replacement. We also denote a *counterfactual sample* as a pair of the counterfactual document and summary, assuming they are factually consistent.

2.1 Formulation

Suppose we have a sample $X_o = (D_o, S_o)$ which consists of document $D_o = \{d_1, d_2, ..., d_M\}$ and a reference summary $S_o = \{s_1, s_2, ..., s_T\}$. We denote a pre-trained language model as ψ and a fine-tuned summarization model as ϕ.

To construct a counterfactual sample $X_c = (D_c, S_c)$ from X_o, we first select the **original named entity** E_o which (*i*) exists in both D_o and S_o and (*ii*) contains the parametric knowledge of ψ. We then replace E_o with the **counterfactual named entity** E_c to synthesize X_c which consists of the counterfactual document D_c and factually consistent summary S_c.

We define factual adaptiveness metrics M_{CL} and M_{FC} on two perspectives: **conditional likelihood** and **factual consistency**, respectively. Specifically, we input original and counterfactual documents alternately into the summarization model, measuring two distinct differences: i) the conditional likelihood of original (counterfactual) named entities within the reference summary and ii) the factual consistency between the original (counterfactual) document and the generated summary.

We define M_{CL} as follows:

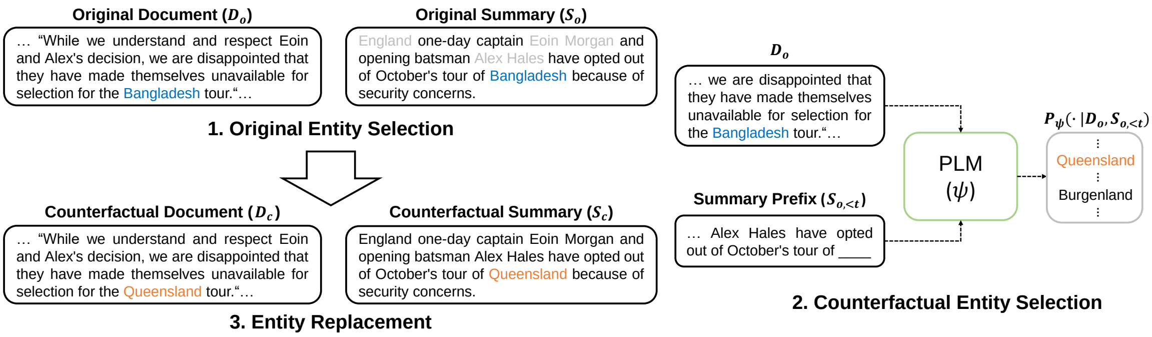
$$M_{CL} := P_\phi(e_o|D_o, S_{o,<t}) - P_\phi(e_c|D_c, S_{c,<t}), \tag{1}$$

where $S_{c,<t}$ and $S_{o,<t}$ denote the summary prefix of first $t-1$ tokens of S_c and S_o, respectively. e_c and e_o denote the first tokens of E_c and E_o, respectively, assuming that e_c and e_o are t-th tokens of each summary. M_{CL} indicates the factual adaptiveness of model ϕ on the perspective of the conditional likelihood when the counterfactual document and the summary prefix are given.

Because M_{CL} does not consider the summary generated by ϕ, we introduce complementary metric M_{FC} as follows:

$$M_{FC} := f(D_o, S^\phi(D_o)) - f(D_c, S^\phi(D_c)), \tag{2}$$

where f denotes factual consistency scoring function such as QuestEval (Scialom et al., 2021), and $S^\phi(D)$ denotes the summary generated by ϕ given


Original Document (D_o)
Original Summary (S_o)
1. Original Entity Selection
Counterfactual Document (D_c)
Counterfactual Summary (S_c)
3. Entity Replacement
D_o
Summary Prefix ($S_{o,<t}$)
PLM
(ψ)
$P_\psi(\cdot|D_o, S_{o,<t})$
Queensland
Burgenland
2. Counterfactual Entity Selection

… "While we understand and respect Eoin and Alex's decision, we are disappointed that they have made themselves unavailable for selection for the Bangladesh tour."…

England one-day captain Eoin Morgan and opening batsman Alex Hales have opted out of October's tour of Bangladesh because of security concerns.

… "While we understand and respect Eoin and Alex's decision, we are disappointed that they have made themselves unavailable for selection for the Queensland tour."…

England one-day captain Eoin Morgan and opening batsman Alex Hales have opted out of October's tour of Queensland because of security concerns.

… we are disappointed that they have made themselves unavailable for selection for the Bangladesh tour."…

… Alex Hales have opted out of October's tour of ____

Figure 1: Overview of the counterfactual sample construction process. The example is sampled from the XSum validation set.

the document D. The second term of M_{FC} involves inputting documents where a knowledge conflict occurs, leading to the generation of factually inconsistent summaries from the model. As a result, M_{FC} approximates factual adaptiveness by calculating the reduction in the factual consistency of the summarization model due to knowledge conflicts.

In the remaining text, we refer to *factual consistency* as the attribute between the **original** document and the generated summary if further clarification is not provided. Note that factual consistency is different from M_{FC} which measures the *difference* of factual consistency scores between original and counterfactual samples.

2.2 Evaluation Set Construction

To satisfy our assumption: (*i*) E_o contains the parametric knowledge of the model and (*ii*) D_c occurs knowledge conflict, it is critical to select appropriate E_o and E_c. We utilize PLM ψ during the entity selection to accurately construct counterfactual sample X_c.

2.2.1 Counterfactual Entity Candidate Pool

We restrict the candidate entities to those of the same category and found in the training corpus following previous works (Longpre et al., 2021; Rajagopal et al., 2022). We utilize *spaCy* (Honnibal et al., 2020) to construct a candidate pool of counterfactual entities from the named entities in the fine-tuning set.

2.2.2 Original Entity Candidates

For each reference summary S_o, we extract the named entity list $L = \{E_{o,1}, E_{o,2}, ..., E_{o,K}\}$ (if $i < j$, $E_{o,i}$ appears before $E_{o,j}$ in S_o) using *spaCy*. In this work, we exclude numerical categories such as `QUANTITY`, `DATE`, and `TIME` concerning that numerical entities can easily be paraphrased (e.g. `15:00` / `3:00 PM`, `1970s` / `70's`).

Algorithm 1 Entity Validation Scenario (S1)

```
Input: Document $D_o = \{d_1, d_2, ..., d_M\}$, summary $S_o = \{s_1, s_2, ..., s_T\}$, pre-trained language model $\psi$, null document $D_\varnothing$, threshold $\tau$.
Output: Counterfactual samples $X_c$

$X_c = \{\}$
$E = \{\}$
Get $L = \{E_{o,1}, E_{o,2}, ..., E_{o,K}\}$, the list of named entity which exists in both $D_o$ and $S_o$
for $k \leftarrow 1$ to $K$ do
    $t_k \leftarrow$ the first token position of $E_{o,k}$ in $S_o$
    $E_c \leftarrow$ named entity sampled from one of three groups    ▷ Section 2.2.3
    $p \leftarrow P_\psi(s_{t_k}|D_\varnothing, S_{o,<t_k})$
    if $p > \tau$ then    ▷ Section 2.2.4
        Append $(E_{o,k}, E_c)$ to $E$
    end if
end for
for each pair $(E_o, E_c)$ in $E$ do    ▷ Section 2.2.5
    $D_c \leftarrow$ REPLACE$(D_o, E_o, E_c)$
    $S_c \leftarrow$ REPLACE$(S_o, E_o, E_c)$
    Append $(D_c, S_c)$ to $X_c$
end for
return $X_c$
```

For each named entity $E_{o,k}$, we validate that the entity is part of the parametric knowledge of ψ. We hypothesize two validation scenarios which will be described in Section 2.2.4.

2.2.3 Counterfactual Entity Candidates

We assume that the original named entity $E_{o,k}$ appears in S_o at the position t_k. We sort counterfactual entity candidates by the conditional likelihood of their first token given the document D_o and the prefix of the reference summary $S_{o,<t_k}$.

We divide the counterfactual entity candidates into three groups: **Top** (top 2%-25% entities by the conditional likelihood), **Middle (Mid)** (25%-75%), and **Bottom (Bot)** (75%-100%). Note that we exclude the top 2% entities to ensure counterfactual replacement. Intuitively, the degree of knowledge conflict is expected to be larger in **Bot** compared to **Top**. We select the group and sample counterfactual entity candidate from the group before the validation step.

2.2.4 Original and Counterfactual Entity Validation

We set two scenarios for the entity validation to satisfy the assumptions described in Section 2.2.

Scenario 1 (S1): Unconditional Likelihood We hypothesize that the named entity $E_{o,k}$ whose unconditional likelihood $P_\psi(e_{o,k}|D_\varnothing, S_{o,<t_k})$ surpasses the threshold τ is part of the parametric knowledge of ψ. $e_{o,k}$ denotes the first token of $E_{o,k}$ (i.e., s_{t_k}), and $D_\varnothing$ denotes the *null document* such as ".". Note that after E_o is validated, we do not further examine E_c in Scenario 1. We refer to Algorithm 1 for details.

Scenario 2 (S2): Conditional Likelihood Difference We hypothesize that $E_{o,k}$ and E_c contain parametric knowledge and knowledge conflict, respectively, if the conditional likelihood difference $P_\psi(e_{o,k}|D_o, S_{o,<t_k}) - P_\psi(e_c|D_c, S_{c,<t_k})$ surpasses the threshold τ. Note that the condition in Scenario 2 is directly aligned to M_{CL} in Equation 1 except for the model to be used. The algorithm of Scenario 2 can be found in Appendix B.

2.2.5 Entity Replacement

If the sample X_o has the valid original (counterfactual) entity E_o (E_c), we replace all E_o in D_o and S_o with E_c. After the entity-level replacement, we further conduct the word-level replacement where each word in E_o is replaced with the word in E_c proportionally to its position.

For example, if E_o = `"Daniel Radcliffe"` and E_c = `"Rupert Grint"`, we further replace `"Daniel"` with `"Rupert"` and `"Radcliffe"` with `"Grint"`.

3 Analysis on Models for Improving Factual Consistency

In this section, we analyze summarization models using the evaluation set as described in Section 2.2. Specifically, we measure M_{CL} and M_{FC} of various models that are proposed to improve factual consistency to observe the relation between factual adaptiveness and factual consistency.

3.1 Setup

We measure ROUGE-L (Lin, 2004) and QuestEval score, which is known to be aligned with human judgments (Scialom et al., 2021), on the original test set and M_{CL}/M_{FC} scores on the factual adaptiveness evaluation set. We use three approaches for the baseline: data filtering (Filtering, Nan et al., 2021), contrastive learning (CLIFF, Cao and Wang, 2021), and advanced decoding (Decoding, Wan et al., 2023) and two backbone PLMs: PEGASUS$_{\text{LARGE}}$ (Zhang et al., 2020) and BART$_{\text{LARGE}}$ (Lewis et al., 2020) for the analysis. We also evaluate models that are simply fine-tuned with negative log-likelihood objectives (NLL) for comparison. For the baseline re-implementation, we use *HuggingFace*[1] for PEGASUS based models and *fairseq*[2] for BART based models. Hyperparameters for each baseline can be found in Appendix E.

3.2 Evaluation Set

We use test sets of XSum (Narayan et al., 2018) and CNN/DailyMail (CNNDM, Hermann et al., 2015) to construct factual adaptiveness evaluation sets. We search the threshold τ using validation sets so that the extracted factual adaptiveness evaluation set is about 10% of the original validation set (We use **Top** group and Scenario 1). τ and dataset statistics can be found in Appendix E.

To specify the evaluation set, information on (i) the type of PLM, (ii) the dataset, (iii) the type of counterfactual entity candidate group, and (iv) the type of validation scenario is required. For example, we denote XSum (PEGASUS, Top, S1) as the evaluation set based on the XSum test set using PEGASUS for the PLM, **Top** group for the counterfactual entity candidate group, and scenario 1 for the entity validation.

3.3 Results

Scores of PEGASUS based models are shown in Table 1. Results on BART based models can be found in Appendix D, except for Decoding because the original training code for BART is implemented on *fairseq*, while the code for Decoding is based on *HuggingFace*.

[1] https://github.com/huggingface/transformers

[2] https://github.com/facebookresearch/fairseq

			M_{CL}(S1)(↓)			M_{CL}(S2)(↓)			M_{FC}(S1)(↓)			M_{FC}(S2)(↓)		
Models	R-L	QEval	Top	Mid	Bot	Top	Mid	Bot	Top	Mid	Bot	Top	Mid	Bot
XSum (PEGASUS)														
NLL	**36.36**	32.94	0.552	0.589	0.631	0.718	0.734	0.744	1.79	2.10	1.98	2.14	2.43	2.55
	±0.07	±0.05	±.004	±.002	±.006	±.004	±.003	±.004	±0.08	±0.08	±0.09	±0.10	±0.02	±0.11
Filtering	34.89	33.49	**0.495**	**0.526**	**0.565**	**0.669**	**0.681**	**0.693**	**1.59**	**1.77**	**1.75**	**1.69**	**1.91**	**2.07**
	±0.09	±0.24	±.011	±.015	±.019	±.010	±.009	±.013	±0.00	±0.12	±0.12	±0.17	±0.08	±0.15
Decoding	35.29	**34.11**	-	-	-	-	-	-	1.76	1.90	2.01	2.09	2.40	2.46
	±0.02	±0.02	-	-	-	-	-	-	±0.11	±0.02	±0.14	±0.19	±0.09	±0.10
CLIFF	35.86	33.27	0.547	0.583	0.625	0.713	0.727	0.740	1.83	2.12	2.11	2.10	2.30	2.50
	±0.04	±0.02	±.005	±.004	±.003	±.006	±.005	±.005	±0.09	±0.16	±0.04	±0.04	±0.16	±0.16
CNN/DailyMail (PEGASUS)														
NLL	37.08	51.44	0.243	0.277	0.304	0.444	0.451	0.449	0.49	0.46	0.45	0.53	0.43	0.44
	±0.05	±0.05	±.003	±.001	±.001	±.002	±.001	±.002	±0.14	±0.04	±0.07	±0.19	±0.14	±0.07
Filtering	36.69	51.86	**0.188**	**0.215**	**0.243**	**0.384**	**0.390**	**0.396**	**0.31**	**0.19**	**0.29**	**0.37**	0.46	0.34
	±0.10	±0.03	±.002	±.001	±.002	±.003	±.001	±0.002	±0.07	±0.09	±0.11	±0.01	±0.05	±0.06
Decoding	**37.52**	**52.60**	-	-	-	-	-	-	0.54	0.41	0.48	0.53	**0.31**	0.49
	±0.10	±0.05	-	-	-	-	-	-	±0.16	±0.18	±0.08	±0.09	±0.13	±0.08
CLIFF	37.06	51.45	0.243	0.278	0.302	0.445	0.452	0.450	0.56	0.60	0.62	0.40	0.50	**0.33**
	±0.04	±0.03	±.002	±.003	±.003	±.001	±.000	±.002	±0.15	±0.11	±0.19	±0.02	±0.10	±0.12

Table 1: Mean and standard deviation of ROUGE-L (R-L) and QuestEval (QEval) on original test sets and M_{CL}/M_{FC} scores on factual adaptiveness evaluation sets across 3 seeds.

Entity Validation Scenarios We first observe which of the two entity validation scenarios more effectively generates knowledge conflict. In most cases, it is observed that factual adaptiveness is much degraded for the evaluation sets constructed based on Scenario 2, especially in XSum. The results suggest that through Scenario 2, we can accurately detect prior knowledge of PLM and effectively induce knowledge conflicts compared to Scenario 1. Given the fact that the criterion used in Scenario 2 is similar to Equation 2, and they only differ in terms of the models used, we speculate that fine-tuned models share the knowledge with pre-trained models.

Counterfactual Entity Candidate Groups We can observe that M_{CL} scores tend to increase in the order of **Top**, **Mid**, and **Bot**. Considering that the group is divided based on the conditional likelihood of PLM, the results indicate that our method controls the degree of parametric knowledge and knowledge conflict effectively.

In CNN/DailyMail, the tendency for M_{FC} between the candidate groups is weak compared to XSum even with the consistency in M_{CL}. We speculate that the low abstractiveness of CNN/DailyMail (Dreyer et al., 2023) has improved overall factual adaptiveness with respect to M_{CL} and M_{FC}, resulting in the similarity of M_{FC} between the candidate groups.

Factual Adaptiveness vs. Factual Consistency While Filtering greatly enhances both factual consistency and factual adaptiveness, Decoding and CLIFF show minimal improvements in M_{CL} and M_{FC} scores compared to NLL. The results imply that methods for factual consistency improvement do not necessarily increase robustness against knowledge conflict, and factual consistency is not strongly correlated with factual adaptiveness.

4 Controllable Counterfactual Data Augmentation

4.1 Training Set Construction

We apply the same procedure used for building a factual adaptiveness evaluation set to construct the augmentation set. For each dataset, we use the same threshold τ determined during the corresponding evaluation set construction. We further proceed to sample the obtained augmentation set at a certain ratio ρ of the original training set.

4.2 Incorporation with Contrastive Learning

In recent research, contrastive learning has been applied to enhance factual consistency (Cao and Wang, 2021; Wan and Bansal, 2022). Our method can be integrated with a contrastive learning-based approach if it can map positive/negative summaries to counterfactual documents.

In the context of contrastive learning, we apply previous contrastive learning set construction methods to the counterfactual samples. For CLIFF, we utilize the provided positive/negative summaries by replacing original entities in the summaries with counterfactual entities. For FactPEGASUS (Wan and Bansal, 2022), we feed augmented datasets to

	XSum								CNN/DailyMail							
	PEGASUS				BART				PEGASUS				BART			
	R-L	QEval	M_{CL}	M_{FC}	R-L	QEval	M_{CL}	M_{FC}	R-L	QEval	M_{CL}	M_{FC}	R-L	QEval	M_{CL}	M_{FC}
NLL	**36.36**	32.94	0.734	2.43	**34.83**	32.94	0.752	2.14	**37.08**	51.44	0.451	0.43	**38.05**	50.99	0.438	0.62
	±0.07	±0.05	±.003	±0.02	±0.05	±0.03	±.004	±0.02	±0.05	±0.05	±.001	±0.14	±0.04	±0.04	±.006	±0.10
CLIFF	35.86	**33.27**	0.727	2.30	33.89	33.32	0.742	2.19	37.06	**51.45**	0.452	0.50	37.97	**51.07**	0.435	**0.52**
	±0.04	±0.02	±.005	±0.16	±0.14	±0.07	±.005	±0.07	±0.04	±0.03	±.000	±0.10	±0.13	±0.06	±.003	±0.10
Ours	35.69	33.26	**0.132**	**1.20**	33.81	**33.39**	**0.113**	**1.20**	36.91	51.37	**0.096**	**0.41**	37.88	51.01	**0.074**	0.56
(CLIFF)	±0.03	±0.06	±.004	±0.10	±0.04	±0.05	±.005	±0.10	±0.00	±0.04	±.001	±0.14	±0.09	±0.03	±.001	±0.21

Table 2: ROUGE-L (R-L) and QuestEval (QEval) on original test sets and M_{CL}/M_{FC} scores on factual adaptiveness evaluation sets of Scenario 2 and **Mid** group with the mean and standard deviation across 3 seeds.

the provided contrastive learning pipelines[3].

In the remaining text, the term **ours** refers to a model that integrates controllable counterfactual data augmentation with the CLIFF training method. We also conduct experiments on FactPEGASUS and experimental results on XSum can be found in Appendix F.

5 Experiments

5.1 Setup

We use Scenario 2 and **Mid** group to construct augmented contrastive learning training sets in accordance with the conclusions drawn in Section 3.3. To regulate the size of the training dataset, we sample the augmentation set from counterfactual samples, setting ρ to 0.1. We use the remaining settings as those of CLIFF in Appendix E. Note that we vary the sampling seed of the counterfactual samples in the multiple seed experiment.

To obtain the positive/negative summaries of the counterfactual document, we utilize the entities E_o and E_c used when obtaining the counterfactual document and apply the same entity replacement process to positive and negative summaries of the corresponding original document. If there is no negative summary for the original document, we obtain it by performing entity replacement on S_o with other counterfactual entities. To gather a sufficient number of negative summaries, multiple counterfactual entity candidates are sampled during the process in Section 2.2.3 before the validation.

5.2 Main Results

We compare the results of our model with those of NLL and CLIFF in Table 2 because CLIFF and ours sequentially apply additional techniques to NLL: contrastive learning and controllable counterfactual data augmentation, respectively.

From the perspective of conditional likelihood (i.e., M_{CL}), we can observe that our method significantly improves factual adaptiveness. Compared to the contrastive learning baseline, our method also enhances factual consistency on the original test set in the BART-XSum case.

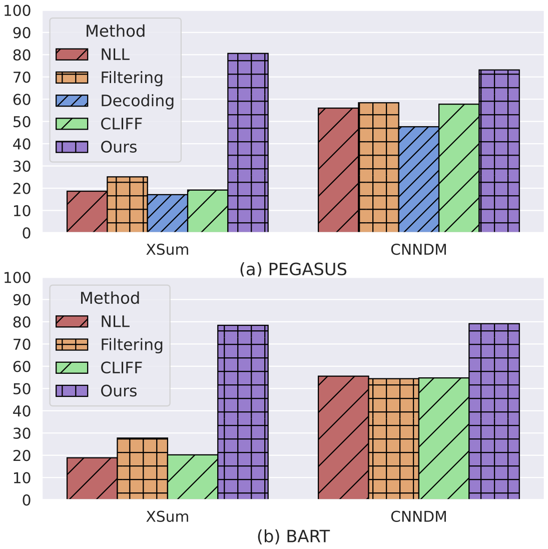


Figure 2: The ratio of summaries generated from the counterfactual documents of XSum and CNN/DailyMail (Mid, S2) which include the counterfactual entity but do not include the original entity.

Although our models consistently reduce the M_{CL} score, there is a case where our M_{FC} score is higher than that of CLIFF in BART fine-tuned with CNN/DailyMail. One possible explanation is that our method is more effective in terms of factual adaptiveness on datasets with a high level of abstractiveness such as XSum, while there is a misalignment between M_{CL} and M_{FC} on datasets with low abstractiveness (Dreyer et al., 2023). We also provide the results of the ChatGPT preference test in Appendix G.

6 Analysis

6.1 Entity Replacement

The proportion of summaries that contain the counterfactual entity without the original entity given

[3] https://github.com/meetdavidwan/factpegasus

	Aug. Group	Aug. Ratio	Evaluation Set			QEval
			Top	Mid	Bot	
XSum	Top	5%	0.191	0.229	0.272	33.35
		10%	**0.147**	0.188	0.236	33.37
	Mid	5%	0.217	0.162	0.156	33.34
		10%	0.163	**0.113**	0.113	**33.39**
	Bot	5%	0.335	0.208	0.127	33.35
		10%	0.285	0.159	**0.081**	33.37
CNNDM	Top	5%	0.142	0.153	0.152	51.04
		10%	**0.102**	0.115	0.118	50.99
	Mid	5%	0.162	0.108	0.078	51.01
		10%	0.118	**0.073**	0.047	51.01
	Bot	5%	0.205	0.124	0.060	**51.08**
		10%	0.173	0.091	**0.028**	50.96

Table 3: Mean of QuestEval (QEval) scores on original test sets and M_{CL} scores on factual adaptiveness evaluation sets of our models based on BART varying the augmentation group (Aug. Group) and augmentation ratio (Aug. Ratio) across 3 seeds.

Dataset	NLL	Filtering	Decoding	CLIFF	Ours
PEGASUS					
XSum	**79.60**	78.08	77.39	78.51	78.26
CNNDM	11.44	9.67	**13.61**	11.23	11.21
BART					
XSum	**80.17**	78.76	-	79.45	79.32
CNNDM	**16.47**	14.42	-	15.99	16.40

Table 4: Mean of MINT scores across 3 seeds.

the counterfactual document is shown in Figure 2.

We can observe that our model exhibits a significantly high rate of generating counterfactual entities in both datasets. Filtering exhibits relatively higher values among the baselines, which is consistent with the results in Table 1. Compared to our method, however, Filtering still generates original entities at a high rate. The results also indicate that our approach successfully addresses entity-level hallucination problems in the BART-CNNDM setting where M_{FC} is slightly higher than that of CLIFF.

6.2 Counterfactual Entity Candidate Group

We vary the counterfactual entity candidate group during the training set construction, as shown in Table 3.

It is observed that M_{CL} scores are minimal when the group type of training set is aligned with the type of evaluation set. We guess that the models tend to fit their factual adaptiveness to the distribution of training sets. It is also observed that models fine-tuned with **Mid** group show low M_{CL} scores across three evaluation sets. Specifically, the score difference between the three evaluation groups of models fine-tuned with **Bot** group is the largest. Based on those observations, we conclude that the distribution of counterfactual samples is important for entity-level generalization of factual adaptiveness.

6.3 Augmentation Ratio

We also vary the augmentation ratio ρ which refers to the ratio of the size of the counterfactual samples to the size of the original training set in Table 3. In all the cases, models of the augmentation ratio of 10% exhibit much lower M_{CL} scores compared to the augmentation ratio of 5%, which implies that the degree of factual adaptiveness can be controlled by modifying ρ. Interestingly, increasing ρ does not always diminish the QEval scores while consistently enhancing factual adaptiveness. The results reemphasize a close-to-orthogonal relationship between factual consistency and factual adaptiveness.

6.4 Factual Adaptiveness vs. Abstractiveness

To observe the relationship between factual adaptiveness and abstractiveness, we measure the MINT abstractiveness score (Dreyer et al., 2023) as shown in Table 4. The abstractiveness of summaries generated by models fine-tuned with XSum demonstrates significantly higher levels of abstractiveness when compared to CNNDM, aligning with the findings of previous studies (Dreyer et al., 2023).

In the baselines, the lowest overall abstractiveness is found in Filtering with the highest factual adaptiveness. On the other hand, our approach demonstrates a relatively minor trade-off between factual adaptiveness and abstractiveness. The results suggest that our method substantially enhances factual adaptiveness while preserving the abstractiveness of generated summaries.

6.5 Qualitative Study

Table 5 shows summarization results given the counterfactual document where the entity `Turkey` is replaced by `Portballintrae`. We use a model weight of $\text{BART}_{\text{LARGE}}$ provided by *HuggingFace*[4] to generate the sample for Decoding. There are clues to infer `Turkey` such as `Kars` and `President Recep Tayyip Erdogan` which result in hallucinated summaries of baselines. On the other hand, our model generates an accurate summary by adapting to the knowledge associated with `Portballintrae`. We present another case study in Appendix H.

[4] https://huggingface.co/facebook/bart-large-xsum

Document (Turkey→Portballintrae)
Ece Heper, 50, was arrested on 30 December in the north-eastern town of Kars, her lawyer Sertac Celikkaleli told The Canadian Press. Canadian officials say they are offering consular assistance, but released no further information. ... Portballintrae's penal code states that anybody who insults the president can face up to four years in prison. … she was arrested for Facebook posts critical of President Recep Tayyip Erdogan. …
Summary
NLL: A Canadian woman has been charged with insulting the president of Turkey, her lawyer says.
Filtering: A Canadian woman has been charged with insulting the president of Turkey, her lawyer says.
Decoding: A Canadian woman has been arrested in Turkey for allegedly insulting the president of the Portballintrae province, her lawyer says.
CLIFF: A Canadian woman has been arrested in Turkey on suspicion of insulting the president, her lawyer says.
Ours: A Canadian woman has been arrested in Portballintrae on suspicion of insulting the president, her lawyer says.

Table 5: Summaries of the counterfactual document of XSum (BART, Mid, S2) evaluation set. Original and counterfactual entities are colored red and blue, respectively.

7 Related Work

7.1 Factual Consistency of Summarization Models

Studies on factual consistency of summarization models have been consistently conducted (Cao and Wang, 2021; Wan and Bansal, 2022; Rajagopal et al., 2022; Wan et al., 2023; Roit et al., 2023). They enhance factual consistency through approaches from various directions such as post-editing (Chen et al., 2021; Balachandran et al., 2022), data augmentation (Rajagopal et al., 2022), contrastive learning (Cao and Wang, 2021; Wan and Bansal, 2022), and advanced decoding (King et al., 2022; Wan et al., 2023).

Rajagopal et al. (2022) synthesize factually inconsistent summaries and augment the corresponding prompts to the document. In this paper, we further modify input documents to trigger knowledge conflict effectively, analyze strategies to consider knowledge conflict, and demonstrate the robustness to entity-level knowledge conflict.

There are also studies focusing on attributes other than factual consistency in summarization models (West et al., 2022; Wu et al., 2022; Cheang et al., 2023). West et al. (2022) analyze whether the model is grounded in the document by ablating facts related to the summary within the document. Wu et al. (2022) analyze the factual robustness, indicating whether the model assigns a low likelihood to an adversarial entity when given the document and factual prompt.

7.2 Parametric Knowledge and Knowledge Conflict

Recent studies in summarization have utilized general-purpose pre-trained language models (Lewis et al., 2020; Raffel et al., 2020; Brown et al., 2020; Ouyang et al., 2022; Chowdhery et al., 2022; Chung et al., 2022) or have pre-trained the language model for summarization (Zhang et al., 2020; Wan and Bansal, 2022).

Recent studies have focused on addressing the hallucination problem in the language model caused by knowledge conflict, especially in question answering domain (Longpre et al., 2021; Neeman et al., 2022; Li et al., 2022; Zhou et al., 2023b).

Ladhak et al. (2023) and Cheang et al. (2023) analyze the hallucination problem of summarization models caused by knowledge conflict in a specific domain: name-nationality knowledge and evolving knowledge over time, respectively. On the other hand, we analyze the robustness of summarization models concerning entity-level knowledge conflicts in arbitrary domains. Moreover, we exploit parametric knowledge from PLM to effectively measure and improve factual adaptiveness.

8 Conclusion

In this study, we analyze the factual adaptiveness of the fine-tuning based summarization models. We propose two complementary metrics of factual adaptiveness and elucidate the relationship between factual consistency and factual adaptiveness. We then propose a controllable counterfactual data augmentation method and observe that our method mitigates hallucination problems due to knowledge conflict. Our experimental results show that our method effectively alleviates entity-level hallucination problems, especially when a knowledge conflict occurs. We anticipate that our work will contribute to improving the faithfulness of summarization models that contain parametric knowledge.

Limitations

In this paper, we conduct entity replacement to synthesize counterfactual samples to control knowledge conflict. Because we utilize *spaCy* to categorize named entity types, the performance of our method can vary depending on the accuracy of the tool. We conduct research on PEGASUS and BART, and further investigation is needed regarding factual adaptiveness in large language models. We focus on entity-level factual adaptiveness, and we leave expanding the scope of knowledge conflict as future work. Future work can also consider orthogonal approaches such as decoding strategy, which can be integrated into our method.

Ethical Considerations

We aim to improve the faithfulness of summarization models in terms of hallucination caused by knowledge conflict which is a major concern of (large) language model based approaches. Our evaluation method could be used to diagnose parametric knowledge and factual adaptiveness which enhances the interpretability of the model.

Acknowledgments

The authors are profoundly grateful for the support provided by the researchers at Samsung SDS. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(2022R1A3B1077720 and 2022R1A5A708390811), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [2022-0-00959 and 2021-0-01343: Artificial Intelligence Graduate School Program (Seoul National University)], Samsung SDS Co.,Ltd, and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024.

References

Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. *arXiv preprint arXiv:2309.04269*.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Chi Seng Cheang, Hou Pong Chan, Derek F Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia S Chao. 2023. Temposum: Evaluating the temporal generalization of abstractive summarization. *arXiv preprint arXiv:2305.01951*.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for*

Computational Linguistics: EACL 2023, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. 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.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. *arXiv preprint arXiv:2203.08436*.

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.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

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 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

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.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Dheeraj Rajagopal, Siamak Shakeri, Cicero Nogueira dos Santos, Eduard Hovy, and Chung-Ching Chang. 2022. Counterfactual data augmentation improves factuality of abstractive summarization. *arXiv preprint arXiv:2205.12416*.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. 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.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

Peter West, Chris Quirk, Michel Galley, and Yejin Choi. 2022. Probing factually grounded content transfer with factual ablation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3732–3746, Dublin, Ireland. Association for Computational Linguistics.

Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Ziqiang Cao, Sujian Li, and Hua Wu. 2022. FRSUM: Towards faithful abstractive summarization via enhancing factual robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3640–3654, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023a. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Algorithm 2 Entity Validation Scenario (S2)

Input: Document $D_o = \{d_1, d_2, ..., d_M\}$, reference summary $S_o = \{s_1, s_2, ..., s_T\}$, pre-trained language model ψ, threshold τ.

Output: Counterfactual samples X_c

```
$X_c = \{\}$
Get $L = \{E_{o,1}, E_{o,2}, ..., E_{o,K}\}$, the list of named entity which exists in both $D_o$ and $S_o$
for $k \leftarrow 1$ to $K$ do
    $t_k \leftarrow$ the first token position of $E_{o,k}$ in $S_o$
    $E_c \leftarrow$ named entity sampled from one of three groups    ▷ Section 2.2.3
    $D_c \leftarrow \textsc{Replace}(D_o, E_{o,k}, E_c)$
    $S_c \leftarrow \textsc{Replace}(S_o, E_{o,k}, E_c)$    ▷ Section 2.2.5
    $e_c \leftarrow$ the first token of $E_c$
    $p_o \leftarrow P_\psi(s_{t_k} | D_o, S_{o,<t_k})$
    $p_c \leftarrow P_\psi(e_c | D_c, S_{c,<t_k})$
    if $p_o - p_c > \tau$ then    ▷ Section 2.2.4
        Append $(D_c, S_c)$ to $X_c$
    end if
end for
return $X_c$
```

A Transferability Test

To clarify that the evaluation set construction method exploits parametric knowledge of PLM rather than global features such as word frequency, we additionally measure M_{CL} and M_{FC} on the evaluation set constructed from other PLM. For example, we evaluate BART fine-tuned on XSum (i.e. BART (XSum)) with the evaluation set XSum (PEGASUS, Mid, S2).

The results of the transferability test are shown in Table 6. M_{CL} and M_{FC} scores of misaligned cases in **Bot** group are lower than the aligned counterparts, implying that we also utilize parametric knowledge not only global attributes during the counterfactual sample synthesis.

B Algorithm of Entity Validation Scenario 2

The detailed content of entity validation scenario 2 is presented in Algorithm 2. The key difference with Algorithm 1 is that Algorithm 2 selects original and counterfactual entities, constructs counterfactual samples, and then calculates the conditional likelihood difference.

Dataset	M_{CL}(S2)(↓) Top	Mid	Bot	M_{FC}(S2)(↓) Top	Mid	Bot
→ BART (XSum)						
XSum (BART)	0.762	0.752	0.757	2.09	2.14	2.26
XSum (PEGASUS)	0.691	0.699	0.694	1.68	1.97	2.19
→ BART (CNNDM)						
CNNDM (BART)	0.472	0.438	0.419	0.56	0.62	0.47
CNNDM (PEGASUS)	0.380	0.357	0.327	0.65	0.59	0.47
→ PEGASUS (XSum)						
XSum (PEGASUS)	0.718	0.734	0.744	2.14	2.43	2.55
XSum (BART)	0.742	0.734	0.725	2.24	2.30	2.35
→ PEGASUS (CNNDM)						
CNNDM (PEGASUS)	0.444	0.451	0.449	0.53	0.43	0.44
CNNDM (BART)	0.458	0.424	0.401	0.59	0.50	0.37

Table 6: Factual adaptiveness results (Scenario 2) when the fine-tuned PLM is aligned/misaligned with the model during the evaluation set construction.

Dataset	R-L	QEval	M_{FC}(S1)(↓) Top	Mid	Bot	M_{FC}(S2)(↓) Top	Mid	Bot
ChatGPT								
XSum	20.74	43.54	1.68	1.53	1.55	1.38	1.48	1.51
CNNDM	31.28	47.71	1.84	1.44	1.82	0.94	0.85	1.14
Ours (CLIFF, PEGASUS)								
XSum	35.69	33.26	1.35	1.23	1.32	1.29	1.20	1.38
CNNDM	36.91	51.37	0.41	0.39	0.35	0.38	0.41	0.38

Table 7: ROUGE-L (R-L) and QuestEval (QEval) scores on original test sets, and M_{FC} scores of ChatGPT and ours on factual adaptiveness evaluation sets using PEGASUS. For ours, each score is the average value for 3 seeds.

C Factual Adaptiveness of ChatGPT

Factual adaptiveness evaluation results of ChatGPT are shown in Table 7. We use *gpt-3.5-turbo-0301* for ChatGPT and utilize PEGASUS to construct factual adaptiveness evaluation sets.

Because PLM which is used to construct factual adaptiveness evaluation sets is not aligned, there is no significant trend between the candidate groups in M_{FC} due to the use of different PLM (i.e., PEGASUS) in the construction of factual adaptiveness evaluation sets. We can observe that factual adaptiveness improves as the model size increases, but it is not completely resolved.

D Baseline Analysis on BART

Baseline analysis results on BART based models are shown in Table 8.

We find that BART does not expose parametric knowledge in entity validation scenario 1. Instead, we observe that replacing the *null document* with the masked summary where named entities are replaced with a special [MASK] token reveals the parametric knowledge. However, we do not further explore the optimal scenario for BART in this pa-

			M_{CL}(S2)(↓)			M_{FC}(S2)(↓)		
Models	R-L	QEval	Top	Mid	Bot	Top	Mid	Bot
XSum (BART)								
NLL	**34.83**	32.94	0.762	0.752	0.757	2.09	2.14	2.26
	±0.05	±0.03	±.004	±.004	±.002	±0.14	±0.02	±0.14
Filtering	31.52	**33.44**	**0.690**	**0.685**	**0.678**	**1.50**	**1.35**	**1.46**
	±0.12	±0.15	±.005	±.008	±.015	±0.04	±0.16	±0.08
CLIFF	33.89	33.32	0.748	0.742	0.747	2.18	2.19	2.36
	±0.00	±0.07	±.002	±.005	±.004	±0.19	±0.07	±0.17
CNN/DailyMail (BART)								
NLL	**38.05**	50.99	0.472	0.438	0.419	0.56	0.62	0.47
	±0.04	±0.04	±.005	±.006	±.005	±0.10	±0.10	±0.11
Filtering	37.53	**51.16**	**0.412**	**0.374**	**0.356**	0.57	**0.49**	**0.34**
	±0.25	±0.02	±0.008	±0.011	±0.012	±0.06	±0.09	±0.11
CLIFF	37.97	51.07	0.470	0.435	0.420	**0.53**	0.52	0.42
	±0.13	±0.06	±.004	±.003	±.002	±0.08	±0.10	±0.01

Table 8: ROUGE-L (R-L) and QuestEval (QEval) on original test sets and factual M_{CL}/M_{FC} scores of BART on factual adaptiveness evaluation sets with the mean and standard deviation across 3 seeds.

	PEGASUS				BART			
	XSum		CNN/DailyMail		XSum		CNN/DailyMail	
	Scenario 1	Scenario 2	Scenario 1	Scenario 2	Scenario 1	Scenario 2	Scenario 1	Scenario 2
Threshold τ	0.05	0.7	0.5	0.75	-	0.6	-	0.65
# Evaluation Set (Top)	1,040	1,003	1,082	1,098	-	1,060	-	1,145
# Evaluation Set (Mid)	1,041	1,163	1,079	1,411	-	1,339	-	1,659
# Evaluation Set (Bot)	1,042	1,326	1,077	1,914	-	1,613	-	2,342
# Train Set (Original)	204,045		287,227		204,045		287,227	
# Train Set (Filtered)	74,241		159,519		74,241		159,519	
# Test Set (Original)	11,334		11,490		11,334		11,490	
Learning Rate	1e-04		5e-05		3e-05		3e-05	
# Train Iter. (Filtered)	10k steps		110k steps		5 epochs		5 epochs	
# Train Iter. (Other)	30k steps		210k steps		5 epochs		5 epochs	

Table 9: Hyperparameters and data statistics.

per to provide general characteristics of fine-tuning based summarization models rather than model-specific analysis. In addition, the tendency of increasing M_{CL} scores in the order of **Top**, **Mid**, and **Bot** groups is observed to be low in BART.

E Hyperparameters and Dataset Statistics

Threshold τ for each evaluation set and dataset statistics are shown in Table 9. Note that the size of the evaluation set of three groups is similar in Scenario 1 because we only use E_o during the validation.

Filtering We exclude samples where at least one named entity in the summary does not appear in the document, except named entities of numerical categories.

CLIFF We choose `SysLowCon` setting used by Cao and Wang (2021)[5]. We use the same objective function and learning rates as those used in CLIFF except for the learning rate during the fine-tuning of PEGASUS with CNN/DailyMail; we use the initial learning rate of 5e-05 following Zhang et al. (2020). We set the coefficient of contrastive loss to 1.0 and the batch size to 8 for both datasets. Regarding the maximum number of negative samples, it is set to 5 for the XSum dataset and 4 for the CNN/DailyMail dataset.

Advanced Decoding We apply the method proposed by Wan et al. (2023) to NLL models and follow `Beam + Greedy Lookahead` setup with a beam width 3[6]. For the XSum dataset, we set the maximum output length to 60 and the look-ahead length to 16. For the CNN/DailyMail dataset, we set the maximum output length to 140 and the look-

[5] https://github.com/ShuyangCao/cliff_summ

[6] https://github.com/amazon-science/faithful-summarization-generation

	R-L	QEval	M_{CL}	M_{FC}
FactPEGASUS	**27.06** ±0.04	34.02 ±0.09	0.597 ±.003	1.59 ±0.09
Ours (FactPEGASUS)	26.79 ±0.06	**34.12** ±0.05	**0.149** ±.004	**1.16** ±0.04

Table 10: ROUGE-L (R-L) and QuestEval (QEval) on XSum test set and M_{CL}/M_{FC} scores of on the factual adaptiveness evaluation set of Scenario 2 and **Mid** group with the mean and standard deviation across 3 seeds.

ahead length to 32.

FactPEGASUS We set the weight of contrastive loss to 5.0 and the maximum number of negative samples to 5. We set the learning rate to 3e-05 and the training step to 15k following Wan and Bansal (2022). The batch size is set to 16, considering that the number of fine-tuning iterations in the original paper is half of that in CLIFF.

F Results on FactPEGASUS

We follow hyperparameters in Appendix E. Threshold τ is set to 0.35 for Scenario 2.

As shown in Table 10, our method can be effectively applied to FactPEGASUS as well. Our method also slightly improves factual consistency on the original XSum dataset compared to the baseline.

G ChatGPT Preference Test

Motivated by Zhou et al. (2023a), we conduct a preference test using ChatGPT for the summaries generated by CLIFF and ours. We use test sets of XSum and CNNDM for the experiment. To remove ordering bias, we randomly shuffle the order of summaries of CLIFF and ours.

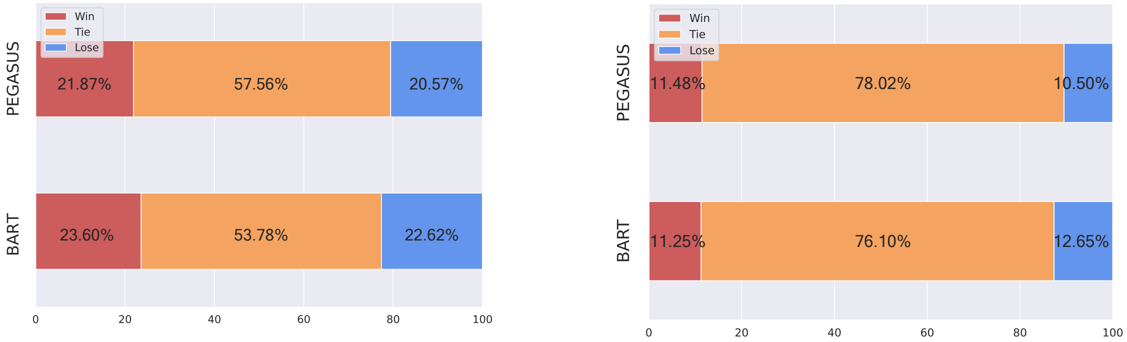
The results are shown in Figure 3. The term *win* indicates that the summary generated by ours is preferred over that of CLIFF. We observe a relatively high proportion of ties in the CNNDM. We speculate that the results are attributed to the low abstractiveness of CNNDM, as mentioned in Section 5.2. When compared to CLIFF, it is observed that ours generally generates preferred summaries for original documents.

H Additional Sample

Other summarization examples are shown in Table 11. Summaries of the baselines generate hallucinated entities instead of reflecting the counterfactual knowledge `Cherry Island`. We speculate that the hallucinations are induced by the relevant entities such as the `UK` and `Northern Ireland`.

I License

The repositories of *fairseq*, FactPEGASUS, and XSum are under the MIT license. The repositories of *HuggingFace*, CLIFF, and CNN/DailyMail are under the Apache-2.0 license. The repository of Decoding is under the CC-BY-NC-4.0, and MINT is under MIT-0.



(a) XSum

(b) CNN/DailyMail

Figure 3: ChatGPT preference test results on (a) XSum and (b) CNN/DailyMail test sets.

Document (London→Cherry Island)
Lagmar Barking, a joint venture between MAR Properties and the Lagan Developments (Holdings), owned the Vicarage Field shopping centre in Barking. It has been bought by Benson Elliot, a UK-based private equity property fund manager. The last accounts for Lagmar Barking showed it owed its lenders £70m. The loan secured on the centre was bought by the US investment fund Cerberus as part of its purchase of the National Asset Management Agency's (Nama) Northern Ireland loan book in 2014. Peter Cornforth, director of retail at Benson Elliot, said the deal was "an exciting purchase for Benson Elliot". He added that it provided the firm with "a fantastic opportunity to contribute to the regeneration of a key east Cherry Island metropolitan centre". Cerberus is continuing to rapidly work through the former Nama portfolio with a combination of asset sales, refinancings and enforcements.
Summary
NLL: A Londonderry shopping centre which was at the centre of a multi-million pound debt has been sold.
Filtering: A Londonderry shopping centre which went into administration last year has been sold for £10m.
Decoding: The former owner of one of Northern Ireland's largest shopping centres has been sold.
CLIFF: A shopping centre in Londonderry has been bought by a private equity firm.
Ours: A shopping centre in east Cherry Island has been bought by a private equity firm for an undisclosed sum.

Table 11: Summarization samples on counterfactual document based on XSum (BART, Mid, S2). Hallucinated entities except for the original named entity E_o are highlighted.