Model Intrinsic Features of Fine-tuning based Text Summarization Models for Factual Consistency

Jongyoon Song[1] **Nohil Park**[1] **Bongkyu Hwang**[2] **Jaewoong Yun**[2]
Seongho Joe[2] **Youngjune L. Gwon**[2] **Sungroh Yoon**[1,3*]
[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
{coms1580, pnoil2588, sryoon}@snu.ac.kr
{bongkyu.hwang, jw0531.yun, drizzle.cho, gyj.gwon}@samsung.com

Abstract

In this study, we analyze the model intrinsic features of a summarization model by varying the fine-tuning objectives and datasets. We fine-tune BART models combining three fine-tuning objectives (negative log-likelihood, unlikelihood, and contrastive loss) and two datasets (CNN/DailyMail and XSum) and provide shuffled or aligned documents to observe changes in the model predictions and intrinsic features. We find that (*i*) the inductive bias for factual consistency during the fine-tuning procedure depends on both the objectives and datasets, and (*ii*) summarization models with relatively low factual consistency are more likely to model summaries that are not conditional to the documents. We demonstrate that splitting data based on the unconditional and conditional summary modeling difficulty affects the factual consistency and intrinsic features of the summarization models. Our experimental results highlight the importance of studying the inductive bias during fine-tuning for factual consistency.

Document
LudoSport has opened its first academy teaching seven forms of combat from the Star Wars world using flexible blades mounted on weighted hilts. The sport began eight years ago in Italy but has only just come to England with the first classes in Cheltenham. Instructor Jordan Court said people were already "hooked". ... So far there are six pupils, but this number is expected to increase. ... The sport is so new to England that there have only been a handful of classes so far ... Lightsaber Combat Academy There are several ranks for those wishing to become a fully-fledged Jedi Knight:
Summary
High $\mathcal{L}_U - \mathcal{L}_C$: A lightsaber-wielding martial artist has opened an academy in Cheltenham to teach people how to "fight like a Jedi".
High $\mathcal{L}_U$: A new martial art inspired by the Star Wars franchise has come to the UK for the first time.
Low $\mathcal{L}_U$: Hundreds of people are taking part in the UK's first lightsaber combat class.
Low $\mathcal{L}_U - \mathcal{L}_C$: If you're a fan of Star Wars, you might want to think twice before taking up lightsaber combat.

Table 1: Output summaries from BART fine-tuned using XSum split in half (high or low) divided by $\mathcal{L}_U$ or $\mathcal{L}_U - \mathcal{L}_C$. $\mathcal{L}_U$ and $\mathcal{L}_C$ indicate unconditional and conditional training losses of summary (see Section 6), respectively. Contents that are not supported by the given document are highlighted.

1 Introduction

Factual consistency in summarization denotes whether the facts in the generated summary are consistent with those of the given document. Factual consistency is essential but remains challenging, particularly in abstractive summarization (Cao et al., 2018; Dong et al., 2020; Huang et al., 2020). Recently, the fine-tuning of pre-trained language models has resulted in an excellent performance with improved factual consistency (Zhang et al., 2020; Cao and Wang, 2021; Wan and Bansal, 2022).

In fine-tuning based summarization, fine-tuning objectives such as contrastive loss (CL) are combined with cross-entropy loss to improve informativeness or factual consistency (Liu and Liu, 2021; Cao and Wang, 2021; Wan and Bansal, 2022). They focus on improving the informativeness and factual consistency aspects of the generated summaries.

Factual consistency also depends on the fine-tuning datasets. In particular, models trained with XSum (Narayan et al., 2018), where the reference summaries are highly abstractive, are known to show degenerated factual consistency (Kryscinski et al., 2020; Maynez et al., 2020; Xie et al., 2021).

There have been several studies on the properties of datasets in this context (Kryscinski et al., 2019; Kang and Hashimoto, 2020; Bommasani and Cardie, 2020; Wan and Bansal, 2022; Liu et al. (2022); Cao and Wang, 2021). For instance, Lin et al. (2022) and Wan and Bansal (2022) report that models trained using XSum are more likely to generate *hallucinated* words, which is critical to the factual consistency. Liu et al. (2022) inspect the dis-

[*] Corresponding author

Findings of the Association for Computational Linguistics: ACL 2023, pages 13884–13898
July 9-14, 2023 ©2023 Association for Computational Linguistics

tribution of predicted hallucinations in terms of the probability and entropy of model predictions. Kang and Hashimoto (2020) propose a loss truncation approach based on the observation that samples with hallucinations have higher losses than other samples. Previous works have focused on dataset intrinsic features or model predictions, but the inductive bias from the datasets and model intrinsic features with respect to the factual consistency has not been fully explored.

In this paper, we propose a unified view of the intrinsic features of a fine-tuning based summarization model by integrating the aspects of fine-tuning objectives and datasets. We utilize BART (Lewis et al., 2020), which has recently been used in fine-tuning based summarization models for factual consistency. We compare three training objectives: negative log-likelihood (NLL), unlikelihood (UL, Welleck et al., 2020; Li et al., 2020), and contrastive loss (Cao and Wang, 2021), and two datasets: CNN/DailyMail (CNNDM, Hermann et al., 2015) and XSum, to fine-tune BART.

We hypothesize that summarization models with low factual consistency are prone to generating summaries less conditioned on the document and perform unconditional summary modeling based on the inductive bias originating from both the training objectives and datasets. To verify our hypothesis, we conduct a *shuffle test*, in which a summary prefix and an aligned (relevant) or shuffled (irrelevant) document are fed to the model to observe the changes in the model intrinsic features. Specifically, we inspect the conditional summary likelihood, summary prefix saliency, and per decoding step entropy during the *shuffle test* to determine the document sensitivity of summarization models with respect to the probability, saliency, and entropy, respectively.

For the dataset aspect, we further hypothesize that the (un)conditional summary modeling difficulty ($\mathcal{L}_U$ and $\mathcal{L}_C$ in Table 1) of data is one of the causes of the inductive bias related to the factual consistency. We fine-tune GPT-2 (Radford et al., 2019) with summaries of the training set using the NLL, then we measure the NLL of each summary as a proxy for unconditional summary modeling difficulty. Similarly, we fine-tune BART with document/summary pairs of the training set using the NLL, then we measure the NLL of each summary as a proxy for conditional summary modeling difficulty. We split the training samples based on the (un)conditional summary modeling difficulty and fine-tune BART models using the subset to inspect the relationship between the (un)conditional summary modeling difficulty of the fine-tuning subset and the factual consistency of the fine-tuned BART.

Based on XSum, we empirically show that a summarization model fine-tuned using a subset with high unconditional and low conditional summary modeling difficulty results in an improved factual consistency relative to a subset with low unconditional and high conditional summary modeling difficulty. Table 1 shows an example summary from the models fine-tuned using different training subsets as split by the (un)conditional training loss. We observe that summaries from models fine-tuned with low $\mathcal{L}_U$ or $\mathcal{L}_U - \mathcal{L}_C$ are factually inconsistent containing errors or unrelated information.

Our findings can be summarized as follows:

- Both the UL and CL based models output decreased summary likelihood and summary prefix saliency given the shuffled document compared to NLL based models.
- BART fine-tuned with XSum is less affected by the information in the documents and tends to unconditionally model the summary.
- In XSum, summaries have a distribution that is easily fine-tuned compared to CNNDM, and samples with low unconditional summary modeling difficulty provide an inductive bias that degrades the factual consistency.

We separate the experiments on the model intrinsic features into those for fine-tuning objectives (Section 5.1) and fine-tuning datasets (Section 5.2) to validate that our methods can be used to diagnose the factual consistency of summarization models. In Section 6, we empirically show that both hallucinations in datasets and unconditional training loss cause an inductive bias that the affects factual consistency.

2 Background

We use BART for all experiments and assume that the summarization model consists of a bidirectional encoder and unidirectional decoder, and that the fine-tuned BART follows an autoregressive decoding scheme. At decoding step t, the summarization model generates a t-th token $\hat{y}_t$ conditioned to document $D = \{d_1, d_2, ..., d_M\}$ where the length is M and the previously decoded sequence $\hat{y}_{<t}$.

Document	Summary	SPS
The space agency has set out a three part plan, which it hopes will eventually lead to humans living on Mars by the 2030s. Unlike the Moon, humans have never physically set foot on Mars, we've only ever used robots like the Curiosity Rover ...	(**Aligned**) Nasa has revealed its plans to try to get humans living on Mars in the next few decades.	0.088
	(**Shuffled**) Microsoft has unveiled Xbox SmartGlass: a service to allow tablet computers and smartphones to communicate with its video games consoles.	0.294
Estimated figures from its Federal Statistical Office said gross domestic product was 1.9% higher last year than in 2015. The annual figure is based on an early estimated ... Household spending grew by 2%, while government spending was up by 4.2%, partly because of an increase in spending ...	(**Aligned**) Germany's economy stepped up its pace of growth in 2016, thanks to higher household and government spending.	0.156
	(**Shuffled**) Police in Edinburgh are investigating a series of thefts and attempted thefts where men have impersonated police officers.	0.488

Table 2: Examples of summary prefix saliency when the document and summary are aligned or shuffled. Note that we shuffle the document-summary pairs in a summary side for readability.

2.1 Fine-tuning Objectives for Factual Consistency

The negative log-likelihood, which is the baseline objective for our experiments, aims to maximize the probability of the reference summary. NLL loss function is defined as follows:

$$\mathcal{L}_{NLL} = -\frac{1}{N}\sum_{n=1}^{N} \log p_n(y_n|y_{<n}, D), \quad (1)$$

where $Y = \{y_1, y_2, ..., y_N\}$ denotes the reference summary where the length is N and p_n denotes the probability distribution at position n. The NLL does not explicitly guide the model to discriminate factually consistent summaries from factually inconsistent summaries.

Recent studies have proposed training objectives such as UL and CL that exploit well-designed, factually inconsistent summaries (negative samples) during training (Cao and Wang, 2021; Wan and Bansal, 2022). The unlikelihood is augmented with NLL to minimize the likelihood of a negative sample $Y' = \{y'_1, y'_2, ..., y'_N\}$ as follows:

$$\mathcal{L}_{UL} = -\frac{1}{N}\sum_{n=1}^{N} \log(1 - p_n(y'_n|y'_{<n}, D)). \quad (2)$$

The contrastive loss in summarization maximizes the similarity between factually consistent and semantically equivalent summaries (positive samples) while minimizing the similarity between positive and negative samples.

The performances of UL and CL highly depend on the discrimination difficulty of the negative samples, and there are various positive and negative sample construction techniques (Kryscinski et al., 2020; Zhang et al., 2021; Cao and Wang, 2021; Wan and Bansal, 2022; Liu et al., 2022).

2.2 Summarization Datasets

We use CNNDM and XSum for the fine-tuning datasets because their characteristics are different. A reference summary in XSum is designed to be more abstractive than a summary in CNNDM. Consequently, multiple studies have shown that models fine-tuned using XSum generate factually inconsistent summaries more frequently than those fine-tuned using CNNDM (Maynez et al., 2020; Xie et al., 2021). Recent studies have shown the presence of *hallucinated* words (i.e. the words that cannot be fully inferred from the given document) in reference summaries lead the model to be factually inconsistent (Lin et al., 2022; Wan and Bansal, 2022).

We also use the Newsroom (Grusky et al., 2018) dataset to compare two BART models fine-tuned using CNNDM and XSum. The reference summary of Newsroom is annotated with the *extractiveness* according to the degree to which words or phrases are included in the document. By utilizing Newsroom, we intend to compare the two models fixing the *extractiveness* and length distribution of the evaluation data as much as possible. We use samples labeled with **Extractive** and **Abstractive** (samples with **Mixed** are excluded) for further analyses.

3 Methodology

We analyze the changes in the model intrinsic features when a document and its summary are factually consistent. We control the factual consistency with document-level perturbation by providing an *aligned* or *shuffled* document-summary pairs to the model.

We choose document-level perturbation to (*i*) maximize the difference in the model intrinsic features and (*ii*) examine whether the model tends to predict a summary conditioned on the document even if the document is irrelevant. In this section, we explain the intrinsic features of the model.

3.1 Conditional Summary Likelihood

Inspired by Xie et al. (2021), we investigate how the (ir)relevant document affects the conditional likelihood of the summary. Given an aligned or shuffled document D and summary y, we define the conditional summary likelihood (CSL) as follows:

$$\mathcal{CSL} = \frac{1}{N}\sum_{n=1}^{N} p_n(y_n|y_{<n}, D). \tag{3}$$

By calculating the difference in the CSL between the *aligned* and *shuffled* cases, we aim to measure the document-sensitiveness of the summarization models when relevant or irrelevant documents are given, respectively.

3.2 Summary Prefix Saliency

We quantify the summary prefix saliency (SPS) to observe whether the model focuses on the document to decode the rest of the summary when the document is irrelevant to the given summary prefix.

The gradient-based input saliency (Li et al., 2016; Sundararajan et al., 2017; Arrieta et al., 2020; Atanasova et al., 2020) is used to explain the prediction in the input space. We aggregate the element-wise multiplication of the input embedding and its gradient using the L2-norm, according to Atanasova et al. (2020).

We first calculate the cross-entropy loss $\mathcal{L}$ from the summarization model ϕ and then derive the gradient with respect to each input embedding. Notably, the loss is calculated using a teacher forcing scheme. We define the saliency of the input token x as follows:

$$\text{Saliency}_x = ||e(x) \odot \nabla_{e(x)}\mathcal{L}_\phi(D, Y)||_2, \tag{4}$$

where $e(x)$ is the embedding vector of the token x and $\odot$ is the element-wise vector multiplication operator.

Based on the input saliency, we calculate the SPS, the input saliency ratio of the summary prefix in the concatenated document and summary, as follows:

$$\mathcal{SPS} = \frac{\sum_{y\in Y} \text{Saliency}_y}{\sum_{x\in D\cup Y} \text{Saliency}_x}. \tag{5}$$

For each target summary token y_n, the SPS quantifies the saliency of the summary prefix $y_{<n}$.

3.3 Per Decoding Step Entropy

Motivated by the study conducted by King et al. (2022) where the entropy of a model is used as a proxy of the uncertainty, we measure the entropy difference between the aligned and shuffled cases. The entropy difference is used to investigate decoding dynamics and approximate the document-sensitiveness, in terms of the uncertainty of the models.

We measure the entropy of the predicted probability distribution at each decoding step n in the teacher forcing scheme as follows:

$$\text{Entropy}_n = -\sum_{i=1}^{|V|} p_n(v_i)\log p_n(v_i), \tag{6}$$

where $V = \{v_1, v_2, ..., v_{|V|}\}$ is the vocabulary of the model. For brevity, we omit the document (D) and summary prefix ($y_{<n}$) conditions in Equation 6. When a shuffled document is given, a relatively low entropy at the decoding step n implies that the model attempts to generate a summarization even if the relevant information does not exist.

By utilizing per decoding step entropy, we focus on analyzing the unconditional summary modeling characteristics of BART fine-tuned with XSum and CNN/DailyMail to investigate inductive bias from the fine-tuning datasets.

4 Experimental Setup

We use BART fine-tuned using the NLL objective provided by *fairseq*[1] (Ott et al., 2019). We refer to the *github* repository of CLIFF[2] (Cao and Wang, 2021) for the fine-tuned models using the CL and scripts for UL based fine-tuning. For the negative sample construction methods, we choose **SysLowCon** for the CL and **MaskEnt** for the UL as proposed by Cao and Wang (2021).

During the CSL and SPS measurements, we calculate the log-probability or cross-entropy loss only for important words. Motivated by Xie et al. (2021), we filter out tokens of specific categories by applying part-of-speech tagging using *spaCy* (Honnibal et al., 2020). We select nouns, verbs, numbers, and proper nouns for the measurements which are more likely to contain factual information. In addition, we filter out the first 30% of the tokens during the loss calculation of the SPS to provide a sufficient summary prefix during the likelihood estimation.

[1] https://github.com/pytorch/fairseq

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

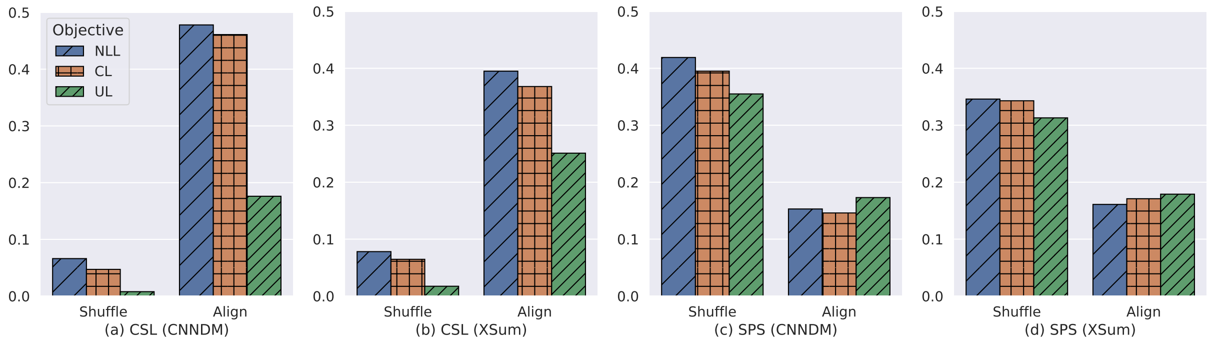


Figure 1: CSL and SPS of BART fine-tuned with three objectives on CNN/DailyMail and XSum test set.

Based on the human evaluation and automatic factual consistency assessment results reported by Cao and Wang (2021), we assume that BART, fine-tuned using CL, UL, and NLL, shows the highest factual consistency in that order, regardless of the fine-tuning datasets.

In the remainder of this paper, we refer to the BART fine-tuned using NLL, UL, and CL as BART-NLL, BART-UL, and BART-CL, respectively. In the same way, we refer to BART fine-tuned using CNNDM and XSum as BART-CNN and BART-XSum, respectively.

5 Fine-tuned Model Analysis

5.1 Fine-tuning Objectives: Conditional Summary Likelihood and Summary Prefix Saliency

Analysis on CSL The results of the CSL and SPS analyses are shown in Figure 1. For both fine-tuning objectives and datasets, the CSL of the shuffled documents is lower than that of the aligned documents.

From a fine-tuning objective perspective, BART-UL and BART-CL show lower CSL values than BART-NLL in the *shuffled* case, and BART-UL shows the lowest CSL values in both the *aligned* and *shuffled* cases. The results imply that factual consistency is negatively related to CSL when an irrelevant document is given; however, preserving a high CSL in *aligned* cases is also needed.

Analysis on SPS In all *shuffled* cases, the SPS is higher than that in *aligned* cases, as shown in Figures 1(c) and (d). This is equivalent to the fact that the model can specify the related context, which is the model intrinsic feature positive to the factual consistency. Both the UL and CL regulate the increment of the SPS in the *shuffled* case from the SPS in the *aligned* case. We conclude that the flexibility of the SPS conditioned on the document is a crucial factor in the factual consistency, and that regulation of the SPS in the *shuffled* case is required for further improvement of the factual consistency.

	Extractive			**Abstractive**		
	Align	Shuffle	Δ	Align	Shuffle	Δ
Negative log-likelihood						
BART-CNN	0.775	0.085	0.690	0.222	0.058	0.164
BART-XSum	0.600	0.083	0.517	0.198	0.057	0.141
Unlikelihood						
BART-CNN	0.339	0.016	0.323	0.101	0.017	0.085
BART-XSum	0.391	0.024	0.368	0.108	0.021	0.086
Contrastive loss						
BART-CNN	0.762	0.068	0.695	0.213	0.046	0.167
BART-XSum	0.549	0.073	0.476	0.187	0.051	0.136

Table 3: CSL of BART-CNN and BART-XSum on Newsroom test set subsets (Extractive, Abstractive).

Comparison between Fine-tuning Datasets We evaluate the CSL of BART-CNN and BART-XSum on the Newsroom dataset to investigate the model intrinsic features caused by the fine-tuning dataset. As shown in Table 3, the CSL differences between the *aligned* and *shuffled* cases of BART-XSum are less than those of BART-CNN regardless of the extractiveness of the summary, except for the unlikelihood objective. We hypothesize that BART-XSum is less conditioned on the information in the document, which is the negative model intrinsic feature for factual consistency. We conduct a further comparison (as discussed in Section 5.2) to validate our hypothesis.

5.2 Fine-tuning Datasets: Per Decoding Step Entropy

We further analyze the dataset aspect because the factual consistency of fine-tuned models largely varies with the dataset characteristics. Figure 2 depicts a graph of the per decoding step entropy averaged over the samples in the Newsroom dataset.

We plot the entropies in the first 100 decoding

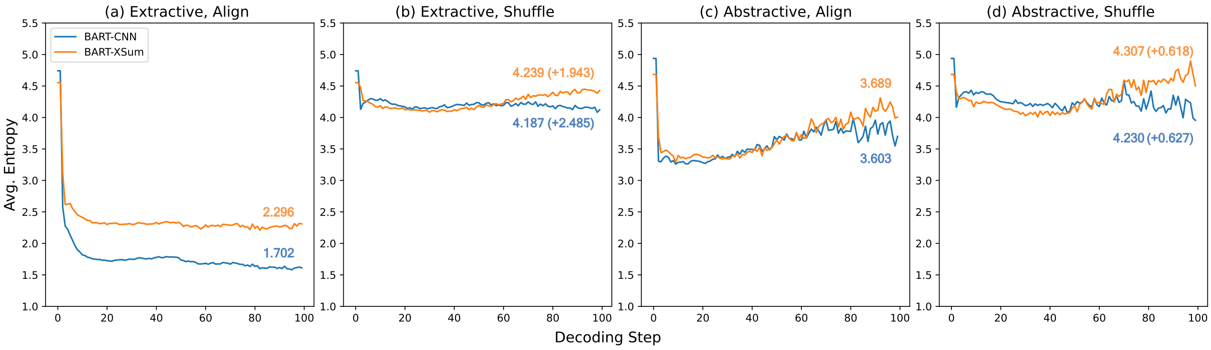
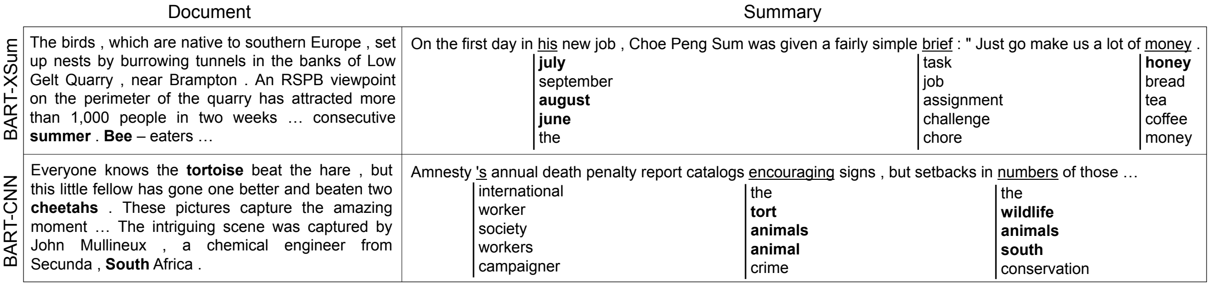


Figure 2: Per decoding step entropy averaged over samples of BART-CNN and BART-XSum on Newsroom test set subsets (Extractive, Abstractive) and two document-summary alignment cases (Align, Shuffle). The entropies averaged over 10–100 decoding steps and their differences between the aligned and shuffled cases are shown on each plot.



	Document	Summary		
BART-XSum	The birds , which are native to southern Europe , set up nests by burrowing tunnels in the banks of Low Gelt Quarry , near Brampton . An RSPB viewpoint on the perimeter of the quarry has attracted more than 1,000 people in two weeks … consecutive **summer** . **Bee** – eaters …	On the first day in his new job , Choe Peng Sum was given a fairly simple brief : " Just go make us a lot of money .		
		july september **august** **june** the	task job assignment challenge chore	**honey** bread tea coffee money
BART-CNN	Everyone knows the **tortoise** beat the hare , but this little fellow has gone one better and beaten two **cheetahs** . These pictures capture the amazing moment … The intriguing scene was captured by John Mullineux , a chemical engineer from Secunda , **South** Africa .	Amnesty 's annual death penalty report catalogs encouraging signs , but setbacks in numbers of those …		
		international worker society workers campaigner	the **tort** **animals** **animal** crime	the **wildlife** **animals** **south** conservation

Figure 3: Examples of top-5 predictions on the decoding steps corresponding to the underlined tokens, given the shuffled document and summary prefix. The emboldened tokens in the document/summary are related to the tokens in the summary/document, respectively.

steps, considering that the maximum summary token length of the XSum test set is approximately 100. Unlike BART-XSum, the average entropy of BART-CNN slightly decreases after 50 decoding steps. One possible explanation is that BART models fine-tuned with longer summaries are able to better predict the next summary tokens. For a quantitative analysis, we calculate the average entropy over the decoding steps 10 to 100. In both the Extractive and Abstractive subsets, BART-CNN shows higher entropy differences between the aligned and shuffled cases, as shown in Figure 2 (b) and (d). Combined with the results in Table 3, we conclude that BART fine-tuned with XSum is less sensitive to a given document.

We hypothesize that the characteristics of the unconditional summary modeling of BART-XSum originate from the inductive bias, which occurs during the fine-tuning phase and degrades the factual consistency of the summarization model. In Section 6, we provide the experimental results regarding the inductive bias of the factual consistency originating from the fine-tuning datasets.

In Figure 3, we visualize the top-5 predictions, assuming that the shuffled document and summary prefix are provided. We gather the top-5 predictions for all the decoding positions simultaneously by following the teacher forcing scheme. First, we identify the cases where the top-5 predictions are strongly conditioned on the summary prefix rather than on the document (e.g., the second and first top-5 predictions in BART-XSum and BART-CNN in Figure 3, respectively). In BART-XSum, there are some predictions weakly related to the words in the document (e.g., summer and bee). However, the predictions are not factually consistent and are not similar to the other predictions (e.g., bread, tea, coffee, and money). In contrast, the top-5 predictions in BART-CNN are more similar to each other or are factually consistent.

6 Modeling Difficulty based Data Splitting and Inductive Bias

To clarify one of the causes of the unconditional summary modeling property of BART-XSum, we fine-tune the GPT-2 model (Radford et al., 2019) using the summaries in the training sets of CNNDM and XSum, and plot the training loss curve

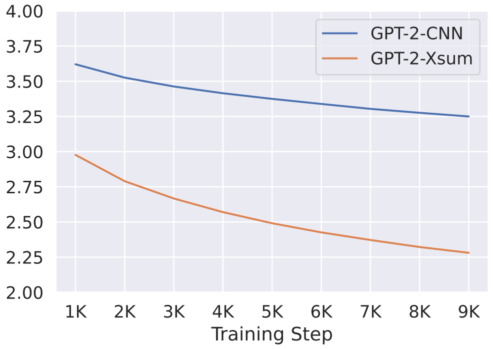


Figure 4: Training loss curves of GPT-2 during fine-tuning using the summaries in CNNDM (GPT-2-CNN) and XSum (GPT-2-XSum).

	XSum		**CNNDM**	
	High	Low	High	Low
$-\mathcal{L}_C$	21.24 ± 0.09	21.05 ± 0.06	46.08 ± 0.24	46.21 ± 0.23
$\mathcal{L}_U$	21.24 ± 0.06	20.86 ± 0.06	46.30 ± 0.15	45.94 ± 0.08
$\mathcal{L}_\Delta$	21.40 ± 0.09	20.77 ± 0.05	46.21 ± 0.24	46.15 ± 0.16

Table 4: QEval scores of BART-XSum and BART-CNN fine-tuned with the subset of training set based on three standards ($-\mathcal{L}_C$, $\mathcal{L}_U$, $\mathcal{L}_\Delta$) and two categories (high, low).

in Figure 4. We use a pre-trained GPT-2 from HuggingFace[3] (Wolf et al., 2020). Based on the results in Figure 4, we conclude that the unconditional language model fine-tuned with the summaries in XSum converges faster than that fine-tuned with CNNDM. This demonstrates our hypothesis in Section 5.2, i.e., that the low unconditional summary modeling difficulty of XSum results in an inductive bias for the unconditional summary modeling property of the fine-tuned summarization model.

6.1 Definition

We use *unconditional* and *conditional training loss* as proxies of unconditional and conditional summary modeling difficulty, respectively. We define the unconditional training loss (denoted as $\mathcal{L}_U$) of the summary as the NLL of the language model fine-tuned with the summary given only the summary. We utilize the fine-tuned GPT-2 to calculate the unconditional training loss. Similarly, we define the conditional training loss (denoted as $\mathcal{L}_C$) of the summary as the NLL of the fine-tuned summarization model given the document and summary. We utilize the fine-tuned BART to calculate the conditional training loss. Additionally, we denote a difference between the unconditional and conditional training loss (i.e. $\mathcal{L}_U - \mathcal{L}_C$) as $\mathcal{L}_\Delta$.

6.2 Setup

We design experiments to observe the relationship between (un)conditional summary modeling difficulty and model intrinsic features.

We split the training set into two subsets based on three standards ($\mathcal{L}_U$, $\mathcal{L}_C$, and $\mathcal{L}_\Delta$), resulting in six fine-tuning subsets in total. We expect that the samples with high $\mathcal{L}_C$ prevent unconditional summary modeling properties and that those with low $\mathcal{L}_C$ assist in the conditional summary modeling.

Our goal is to analyze the six summarization models with respect to factual consistency, CSL, and SPS. For the factual consistency evaluation, we leverage the QuestEval (QEval in short) score (Scialom et al., 2021), a question-answering based automated evaluation method.

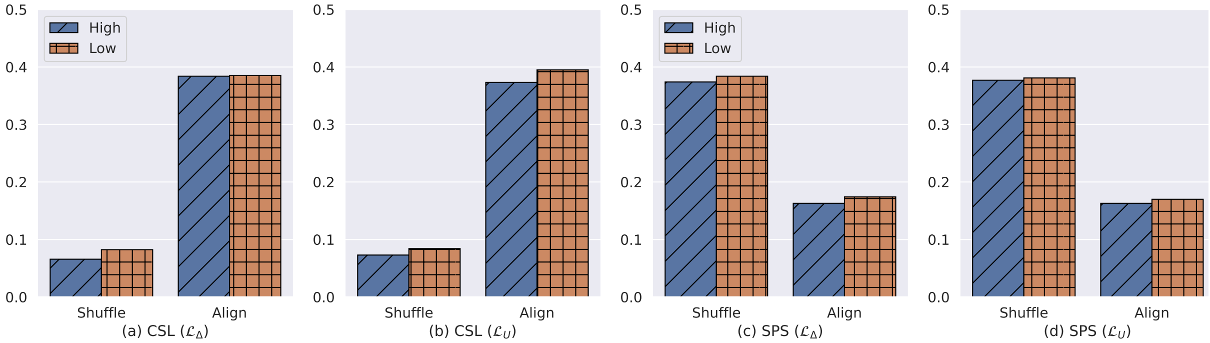
Some studies have controlled training samples during summarization model training to improve the factual consistency (Kang and Hashimoto, 2020; Wan and Bansal, 2022; Goyal and Durrett, 2021). For instance, Kang and Hashimoto (2020) adaptively eliminate samples with high loss and Goyal and Durrett (2021) propose identifying non-factual words in a reference summary and modifying the objective function for the factual consistency. Unlike previous works, we fine-tune both conditional and unconditional language models to calculate the difficulty and analyze the inductive bias during fine-tuning.

6.3 Results

Factual Consistency Analysis We measure the QEval scores of BART fine-tuned using the six categories of training subsets. Table 4 presents the means (with a 99% confidence interval) of 8 different seeds.

In XSum, subsets with low $\mathcal{L}_C$ or high $\mathcal{L}_U$ achieve higher factual consistency compared to their counterparts. The difference in the QEval scores between the *high* and *low* subsets of $\mathcal{L}_U$ is larger than those of $\mathcal{L}_C$, implying that a low $\mathcal{L}_U$ causes the inductive bias of the factual consistency degradation in BART-XSum. Furthermore, collecting samples with a high $\mathcal{L}_\Delta$ maximizes the factual consistency. Documents with high $\mathcal{L}_\Delta$ vastly decrease the conditional summary modeling difficulty by using the related information in the document; simultaneously, they roughly filter out the hallucinated samples.

[3] https://huggingface.co/gpt2



(a) CSL ($\mathcal{L}_\Delta$)

(b) CSL ($\mathcal{L}_U$)

(c) SPS ($\mathcal{L}_\Delta$)

(d) SPS ($\mathcal{L}_U$)

Figure 5: CSL and SPS of BART fine-tuned with the subset of XSum training set based on two standards considering unconditional training loss ($\mathcal{L}_U$, $\mathcal{L}_\Delta$).

In CNNDM, data splitting based on $\mathcal{L}_C$ and $\mathcal{L}_U$ affects the factual consistency less, than that in the case of the XSum results. One possible explanation is that the negative inductive bias for the factual consistency from the dataset is less dominant than the inductive bias from other factors, such as the fine-tuning objectives.

Model Intrinsic Feature Analysis We further analyze BART-XSum with regards to the CSL and SPS to determine the correlation between $\mathcal{L}_U$ and resulting model intrinsic features. In Figure 5, we visualize the CSL and SPS based on $\mathcal{L}_U$ and $\mathcal{L}_\Delta$, because we assume that $\mathcal{L}_U$ is the major factor in the unconditional summary modeling characteristics of the models fine-tuned using XSum.

It is observed that the models fine-tuned with XSum subsets of the *high* group show lower CSL and SPS in the *shuffle* case than the models of the *low* group. When comparing $\mathcal{L}_U$ and $\mathcal{L}_\Delta$, it is observed that the gap in the CSL between the models in the *high* and *low* groups is larger in $\mathcal{L}_\Delta$. The regulated CSL and SPS of the models of the *high* group in the *shuffle* case indicate that $\mathcal{L}_U$ is related to the inductive bias of unconditional summary modeling and factual consistency.

7 Related Work

7.1 Fine-tuning Methods for Factual Consistency Enhancement

Sequence-to-sequence models, widely used in text generation tasks such as language modeling (Devlin et al., 2019; Radford et al., 2019) and summarization (Lewis et al., 2020; Zhang et al., 2020) are commonly trained under the NLL.

Despite their powerful modeling performance, language models trained to optimize the NLL often encounter the text degeneration problem where the generated texts contain repetitive words or inconsistent contexts (Holtzman et al., 2020). UL was proposed to prevent such problems by penalizing the modeling probability of unwanted tokens (Welleck et al., 2020; Li et al., 2020).

CL was also proposed to enhance factual consistency in the summarization models. Cao and Wang (2021) fine-tune pre-trained language models with a contrastive loss to learn distinguishable representations of the factually erroneous summarization outputs from the sound ones. Wan and Bansal (2022) propose factual consistency enhancing methods with pre-training objectives and fine-tuning modules.

In this paper, we compare the fine-tuning objectives for factual consistency with respect to the conditional summary likelihood and summary prefix saliency to find the relationship between the fine-tuning objectives and the model intrinsic features.

7.2 Analysis of Summarization Datasets and Models

There have been several studies analyzing the properties of summarization datasets (Kryscinski et al., 2019; Maynez et al., 2020; Bommasani and Cardie, 2020). For instance, Maynez et al. (2020) inspect intrinsic/extrinsic hallucinations in model-generated summaries on the XSum dataset, and Bommasani and Cardie (2020) quantify the intrinsic features of datasets, such as their topic similarity and abstractivity. In this paper, we focus on the inductive bias from the datasets and resultant model intrinsic features rather than the characteristics of the dataset itself.

On the other hand, the characteristics of the abstractive summarization model have also been studied (Kang and Hashimoto, 2020; Pagnoni et al., 2021; Xu et al., 2020; Cao and Wang, 2021; Liu et al., 2022; West et al., 2022). Xu et al. (2020) relate model intrinsic features such as entropy and attention to a model prediction. Liu et al. (2022) and Cao and Wang (2021) analyze the characteristics of the predicted hallucinated words with respect to model intrinsic features such as probability and entropy. West et al. (2022) conduct factual ablation studies to observe the probability differences when essential information is ablated from a document.

In this paper, we analyze the model with respect to the datasets and fine-tuning objectives and focus on the intrinsic features of the fine-tuned model, such as the SPS and unconditional summary modeling property. Unlike West et al. (2022), we provide aligned and shuffled documents to maximize the CSL, SPS, and entropy differences.

Kang and Hashimoto (2020) report that the losses of samples containing hallucinations are higher than those of others. We also exploit the loss during the data splitting experiments, but we additionally utilize the unconditional modeling loss of the summary. Furthermore, we integrate the loss with the inductive bias of the fine-tuned models.

8 Conclusion

In this work, we analyze the model intrinsic features that contribute to factual consistency enhancement of the summarization model. With the assumption that summarization models with low factual consistency tend to ignore information from the given documents, we conduct shuffle tests and propose a unified view of the intrinsic features on the fine-tuning objectives and datasets. We measure the CSL, SPS, and per decoding step entropy of BART models to clarify the model intrinsic features related to factual consistency. We also correlate the unconditional and conditional summary modeling difficulty with the inductive bias for factual consistency through data splitting experiments. Based on the analyses, we anticipate that our method can be used as an indicator of the factual consistency of summarization models.

Limitations

In this paper, we use a pre-trained BART to observe the changes in the model intrinsic features by varying the fine-tuning objectives and datasets. However, our methods can be made significantly more generalizable when the range of summarization models (e.g., PEGASUS (Zhang et al., 2020)) and datasets (e.g., FRANK benchmark (Pagnoni et al., 2021)) are broadened. Additionally, we can try using other evaluation methods, such as FactCC (Kryscinski et al., 2020), reported to have a high correlation with human judgment (Pagnoni et al., 2021) to interpret the model intrinsic behavior during the *shuffle tests*.

We leave the research on methods to optimize the inductive bias during the fine-tuning process to improve the factual consistency as future work. Our experimental results can also be integrated with those of previous studies focusing on hallucinations in datasets (Kryscinski et al., 2020; Wan and Bansal, 2022).

Acknowledgement

The authors would like to thank the researchers at Samsung SDS for their invaluable support. This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (2022R1A3B1077720 and 2022R1A5A708390811), Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) [2022-0-00959 and 2021-0-01343: Artificial Intelligence Graduate School Program (Seoul National University)], Samsung Electronics Co., Ltd (IO221213-04119-01), SNU-Naver Hyperscale AI Center, and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023.

Ethics Statement

Our goal is to diagnose factual consistency of fine-tuning based summarization models and prevent models from providing human with misleading information. By connecting model intrinsic features and factual consistency, we aim to improve explainability and faithfulness of summarization models. One of major concerns is that there are still aspects of the model which are not explored, such as scaling law of factual consistency. Continuous studies on model intrinsic features would result trustful summarization models which we leave as future work.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, 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.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

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

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

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, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

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.

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.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. 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 1065–1072, Online. Association for Computational Linguistics.

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.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

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

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.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

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.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Sen Zhang, Jianwei Niu, and Chuyuan Wei. 2021. Fine-grained factual consistency assessment for abstractive summarization models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 107–116, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

	Train	Validation	Test
XSum	204,045	11,332	11,334
CNN/DailyMail	287,227	13,368	11,490
Newsroom (Ext)	332,131	36,372	36,165
Newsroom (Abs)	333,350	36,480	36,595

Table 5: Dataset statistics of XSum, CNN/DailyMail, Newsroom of Abstractive summary (Abs), and Newsroom of Extractive summary (Ext).

	FT Steps	FT Time	GPU(s)
BART-CNN	20k	6 hours	Tesla V100×4
BART-XSum	15k	4 hours	Tesla V100×4
GPT-2-CNN	12k	9 hours	Tesla V100×2
GPT-2-XSum	9k	7 hours	Tesla V100×2

Table 6: Fine-tuning (FT) steps, time, and GPUs for the models used in our experiments.

A Fine-tuning Language Model for Dataset Split

To split the fine-tuning datasets based on the modeling difficulties, we fine-tune the pre-trained GPT-2-base model (Radford et al., 2019) from HuggingFace[4] (Wolf et al., 2020) using the reference summary set from each dataset. We re-train the byte-pair encoding tokenizer of the GPT-2 model with XSum (or CNN/DailyMail) subset for better adaptation on the summarization domain. The final language models using CNN/DailyMail and XSum are fine-tuned for 12,000 and 9,000 steps, respectively, using early stopping with the validation loss. We split the fine-tuning dataset to a block size of 1024, use a batch size of 32, and set an initial learning rate of 5e-5.

B Decoding Details

We follow the same decoding scheme, i.e., beam search decoding, as Cao and Wang (2021). We also use the same hyperparameters of decoding: *fairseq*[5] (Ott et al., 2019) and CLIFF[6] (Cao and Wang, 2021).

For CNN/DailyMail, we set the beam width to 4 and the minimum length to 55. For XSum, we set the beam width to 6 and the minimum length to 10.

	NLL	CL
CNNDM (Align)	0.4485	0.3382
CNNDM (Shuffle)	0.0819	0.0354
XSum (Align)	0.4239	0.3779
XSum (Shuffle)	0.1076	0.0885

Table 7: CSL of PEGASUS fine-tuned with NLL and CL objectives on CNN/DailyMail and XSum test set.

C Dataset Statistics

Statistics of XSum, CNN/DailyMail, and Newsroom are shown in Table 5. Note that we only use the test set of Newsroom because we use Newsroom during the evaluation of BART-CNN and BART-XSum.

D Model Size and Training Time

We fine-tune BART-large which consists of 400M parameters for the experiments. To calculate unconditional training loss of summary, we fine-tune GPT-2 which consists of 117M parameters. Detailed fine-tuning times for the models are shown in Table 6.

E License of Repositories and Datasets

The repository of *fairseq* is under the MIT license, and the repository of *huggingface* is under the Apache-2.0 license. The repositories of CNN/DailyMail and Newsroom are under the Apache-2.0 license, and the repository of XSum is under the MIT license.

F Conditional Summary Likelihood on PEGASUS

We compare the CSL of fine-tuned PEGASUS (Zhang et al., 2020) with respect to fine-tuning objective following hyperparameters as Cao and Wang (2021). Table 7 shows results similar to BART: in the *shuffle* case, PEGASUS fine-tuned using the CL shows less CSL compared to its NLL counterpart.

[4] https://huggingface.co/gpt2

[5] https://github.com/facebookresearch/fairseq/blob/main/examples/bart/summarize.py

[6] https://github.com/ShuyangCao/cliff_summ/tree/main/scripts/bart

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
"Limitations", just after "8 Conclusion" (without section number)

☒ A2. Did you discuss any potential risks of your work?
Analyzing and enhancing factual consistency of text summarization will reduce potential risk of fake news or false information.

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

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

"5 Fine-tuned Model Analysis", "6 Modeling Difficulty based Data Splitting and Inductive Bias"

☑ B1. Did you cite the creators of artifacts you used?
"2.2 Summarization Datasets", "4 Experimental Setup", "5 Fine-tuned Model Analysis", "6 Modeling Difficulty based Data Splitting and Inductive Bias"

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
"Appendix E License of Repositories and Datasets"

☒ 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)?
Provider of used artifacts did not provide intended use.

☒ 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?
Provider of used artifacts did not provide such steps.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We provide examples of datasets which indirectly provide domain and language information.

☑ 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 C Dataset Statistics"

C ☑ Did you run computational experiments?

"5 Fine-tuned Model Analysis", "6 Modeling Difficulty based Data Splitting and Inductive Bias"

☑ 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 A Fine-tuning Language Model for Dataset Split", "Appendix D Model Size and Training Time"

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?
We follow hyperparameters of previous work. For GPT-2 fine-tuning, we did not conduct hyperparameter search because we observe that training/validation loss is converged , and we use GPT-2 for the comparison rather than the absolute value evaluation.

☑ 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?
"6.3 Results"

☒ 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.)?
We use default settings and hyperparameters of Spacy and QuestEval.

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

Left blank.

☐ 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.?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.