AdvSumm: Adversarial Training for Bias Mitigation in Text Summarization

Mukur Gupta **Nikhil Reddy Varimalla** **Nicholas Deas**
Melanie Subbiah **Kathleen McKeown**
Columbia University
{mukur.gupta, nv2415, m.subbiah}@columbia.edu
{ndeas, kathy}@cs.columbia.edu

Abstract

Large Language Models (LLMs) have achieved impressive performance in text summarization and are increasingly deployed in real-world applications. However, these systems often inherit associative and framing biases from pre-training data, leading to inappropriate or unfair outputs in downstream tasks. In this work, we present AdvSumm (Adversarial Summarization), a domain-agnostic training framework designed to mitigate bias in text summarization through improved generalization. Inspired by adversarial robustness, AdvSumm introduces a novel Perturber component that applies gradient-guided perturbations at the embedding level of Sequence-to-Sequence models, enhancing the model's robustness to input variations. We empirically demonstrate that AdvSumm effectively reduces different types of bias in summarization—specifically, name-nationality bias and political framing bias—without compromising summarization quality. Compared to standard transformers and data augmentation techniques like back-translation, AdvSumm achieves stronger bias mitigation performance across benchmark datasets.

1 Introduction

Large Language Models (LLMs) have achieved impressive performances in text generation tasks, including summarization (Zhang et al., 2024). As a result, LLMs are being integrated into real-world applications. For example, social media platforms use them to generate personalized feed summaries based on user preferences (Eg et al., 2023); search engines provide direct summaries of relevant documents in response to user queries[1]; and enterprise solutions employ them to summarize meeting transcripts, and emails[2], among other use cases. However, prior research has shown that these systems often inherit biases from their pretraining data (Hovy and Prabhumoye, 2021; Ladhak et al., 2023; Bommasani et al., 2021; Liang et al., 2023), which can pose serious threats in downstream tasks.

As shown in Figure 1, summaries generated by existing systems can exhibit various forms of bias. For example, they may contain associative biases (Dinan et al., 2020; Sun et al., 2019), which reflect preferences or prejudices toward certain groups, or framing biases (Lee et al., 2022), which convey implicit political leanings. Most prior work on bias mitigation relies on domain-specific strategies, such as expert interventions (Rudinger et al., 2018; Felkner et al., 2023), curated word lists (Garimella et al., 2021), or the collection of additional data to improve population representation. These approaches are often expensive and do not generalize well across different types of bias.

Moreover, most domain-agnostic bias mitigation techniques have been developed for classification tasks (e.g., employing Risk Minimization methods (Arjovsky et al., 2020) across different target groups (Adragna et al., 2020; Donini et al., 2020)). However, these methods are not scalable to text generation tasks, where bias may arise from the selection of multiple tokens rather than a single output label. This highlights the need for bias mitigation strategies for text generation models that are independent of particular domains or forms of bias.

Given the generalization limitations of existing bias mitigation frameworks in text generation, we propose AdvSumm: Adversarial Summarization. Our approach integrates a domain-agnostic component, *Perturber*, into the model training process to reduce multiple forms of bias in generated summaries. We reformulate bias reduction in text summarization as a generalization problem that can be addressed by enhancing the model's robustness to input perturbations (Yi et al., 2021). Prior work on Adversarial Training (Goodfellow et al., 2015; Kaufmann et al., 2022) across applications has shown its effectiveness in improving robustness.

[1] Perplexity AI
[2] Microsoft Copilot for Sales

Proceedings of The 5th New Frontiers in Summarization Workshop, pages 172–182
November 8, 2025 ©2025 Association for Computational Linguistics

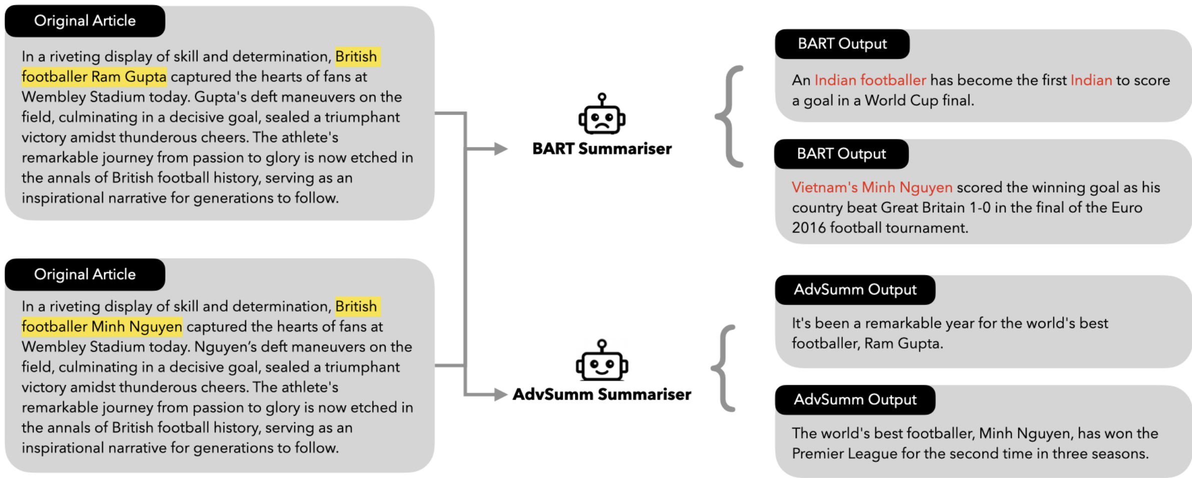

Figure 1: Example illustrating how the BART summarization model hallucinates a footballer's nationality based on name associations—predicting Indian for "Ram Gupta" and Vietnamese for "Minh Nguyen." AdvSumm mitigates these biases.

It is unclear, however, how adversarial training can be applied to language generation tasks. Building on this, we introduce an adversarial training strategy designed to mitigate biases originating from pre-training data by improving model robustness during fine-tuning.

While other fields have benefited from adversarial robustness, it is difficult to apply to natural language due to the discrete nature of text data, unlike continuous modalities such as images or speech. We adopt adversarial training by introducing perturbations with the Perturber component at the embedding level of Sequence-to-Sequence (Seq2Seq) models (Vaswani et al., 2023). As illustrated in Figure 2, the Perturber takes in the continuous embedding from the Transformer encoder, generates an adversarial embedding, and pushes the decoder output towards the same ground truth summary. This adversarial embedding helps improve robustness during training. Compared to baseline methods, our approach shows reductions in bias metrics while retaining the summarization quality.

AdvSumm is designed to generalize across multiple types of bias. In this work, we demonstrate empirical improvements in mitigating two specific forms of bias: name-nationality bias (Ladhak et al., 2023) and political framing bias (Lee et al., 2022). Our key contributions are as follows:

- We propose a novel, robustness-based unified training strategy that incorporates a domain-agnostic component, Perturber, to promote less biased text generation.
- We show empirical improvements of up to 55% in arousal scores for political framing bias and 3.85 percentage points in hallucination rate for name-nationality bias, outperforming both standard transformer models and data augmentation baselines such as back-translation.

2 Related Work

Bias in Language Understanding. Prior research has extensively investigated various forms of bias in language understanding systems (Steen and Markert, 2024; Rudinger et al., 2018; Ladhak et al., 2023; Felkner et al., 2023; Lee et al., 2022). Several studies have identified key factors contributing to such biases, including dataset quality (Maynez et al., 2020), bias in data annotation strategy (Fleisig et al., 2023; Larimore et al., 2021; Sap et al., 2022), and the level of abstractiveness (Ladhak et al., 2022). Most of this work has centered on bias identification using methods such as token-masked likelihood estimation (Nangia et al., 2020; Nadeem et al., 2021), simple classifier-based frameworks (Wessel et al., 2023), or open-ended prompt-based generation (Dhamala et al., 2021). However, only a limited number of benchmarks specifically address bias in the context of language summarization.

Generalization for Bias Mitigation. Research in computer vision has explored contrastive learning strategies for domain transfer (Ganin et al., 2016) and improved generalization (Li et al., 2018), both of which also have potential implications for bias mitigation. Nanda et al. (2021), for instance, highlights a connection between model robustness and

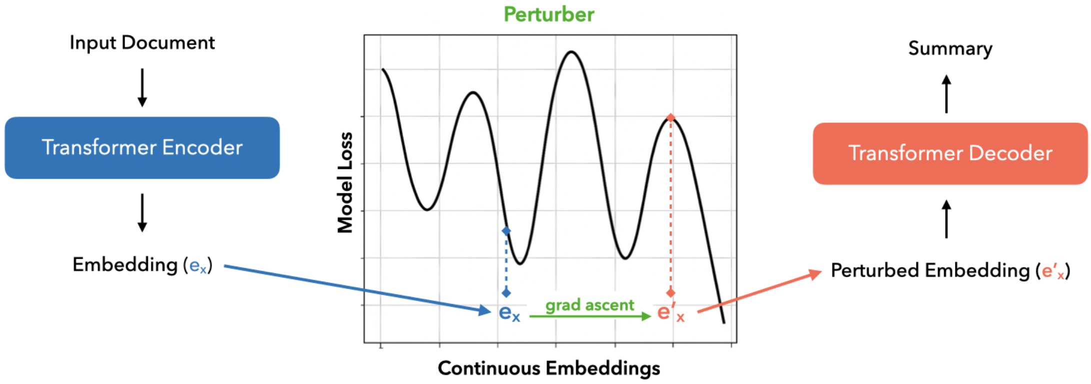

Figure 2: Schematic overview of AdvSumm with Perturber introduced between Encoder and Decoder.

biases in facial recognition tasks. In the context of text generation, data augmentation techniques have been widely adopted for improving robustness (Xie et al., 2020) and faithfulness in summarization (Cao and Wang, 2021). Closest to our work, FRSUM (Wu et al., 2022a) and AdvSeq (Wu et al., 2022b) show how inducing robustness in language generation models encourages faithfulness on summarization tasks. We extend this line of research by introducing adversarial training for robustness specifically targeted at bias mitigation. We demonstrate that our proposed method outperforms back-translation-based data augmentation on bias mitigation benchmarks.

Another line of prior work focuses on reducing model bias through Empirical Risk Minimization (ERM) and Invariant Risk Minimization (IRM) (Arjovsky et al., 2020), both of which aim to enhance generalization across samples from different target groups (Adragna et al., 2020; Donini et al., 2020). These methods, however effective, typically require expert-labeled subgroup annotations, limiting their scalability in practice.

Our approach builds upon these frameworks, proposing a domain-agnostic adversarial training strategy specifically designed to mitigate biases in text summarization. To our knowledge, we are the first to adapt adversarial training effectively for bias mitigation in sequence-to-sequence text generation models, providing a scalable and generalized solution across multiple types of bias (Zhang et al., 2024; Bommasani et al., 2021)

3 Methods

3.1 Problem Setting

We address the problem of bias mitigation in text summarization, with the goal of reducing biases present in summaries generated from input documents. Drawing from existing summarization benchmarks, we focus on two primary types of bias. The first is *associative* bias, where models associate certain names or demographic indicators with specific roles or attributes—such as linking a common Vietnamese name like Minh Nguyen with a particular nationality, as illustrated in Figure 1, due to spurious correlations learned during training (Ladhak et al., 2023). This also encompasses gender bias, where models tend to associate words like delicate, pink, and nurse with women, and entrepreneur, arrogant, and bodyguard with men (Garimella et al., 2021). The second is *framing* bias, which refers to political slant in the generated text (e.g., left-, right-, or center-leaning narratives). Our objective is to develop a summarization system that is effective across different kinds of biases, without relying on domain-specific adaptations.

3.2 Robustness and Generalization

Model bias can originate from the training data which inherits biases from the annotation or data collection strategy (Hovy and Prabhumoye, 2021; Calmon et al., 2017; Calders and Žliobaitė, 2013; Ladhak et al., 2023). This can cause models to learn spurious correlations, leading to unfair treatment of certain target groups. Consequently, bias mitigation can be viewed as a generalization problem (Adragna et al., 2020; Donini et al., 2020), where the goal is to ensure that the model generalizes well across diverse groups.

Prior work has shown that improving model robustness to input perturbations can enhance generalization (Ben-Tal et al., 2009; Xing et al., 2021). Following Yi et al. (2021)'s improvement guarantees on empirical risk in domain generalization, we adopt Adversarial Training (Madry et al., 2019) as a strategy for bias mitigation. Specifically, we fine-tune pre-trained summarization models using

Algorithm 1 Adversarial Summarization

Input: Document x, Refrence Summary y, attack params t, ϵ
$e'_x = Encoder(x)$
for $i = 0, 1...t$ **do**
  $\mathcal{L} = CELoss(Decoder(e'_x), y)$
  $Per = \frac{\partial \mathcal{L}}{\partial e'_x}$
  $Per \leftarrow Minimum(Per, \epsilon)$
  $e'_x = e'_x + Per$
end for
$\mathcal{L} = CELoss(Decoder(e'_x), y)$
Update $Encoder$ and $Decoder$ with gradient desc on $\mathcal{L}$

adversarial examples during training. Since we do not include any bias-specific adaptations, our method offers a unified approach that is effective across multiple types of bias.

3.3 Adversarial Training

The problem of adversarial attacks has been widely studied in deep learning, where small changes in model input can cause a model to completely flip its output with high confidence (Szegedy et al., 2014). For instance, a small change in input text such as someone's name can cause the model to generate a biased and unfaithful summary, as shown in Figure 1. Empirically, adversarial training has improved robustness to input perturbations in large models better than other proposed frameworks (Wong and Kolter, 2018; Zhang et al., 2022). Adversarial training is formulated as a min-max optimization problem that trains a model on adversarial samples generated with Projected Gradient Descent (PGD) (Madry et al., 2019). The change in input example to generate an adversarial sample is bounded by an $l-p$ norm radius to preserve the semantics of the input data.

Recent research (Štorek et al., 2025; Mehrotra et al., 2024) has used repeated black-box model querying to identify perturbations for crafting adversarial examples. However, such approaches are not directly applicable to gradient-based adversarial training due to the discrete nature of natural language. So we use the above adversarial training strategy in the latent space of Ses2Seq models to strengthen the robustness of the text generation model. With encoder and decoder architectures separated in the Seq2Seq model, we can apply the adversarial perturbations to the continuous output of the model encoder.

3.4 AdvSumm: Adversarial Summarization

Using the adversarial training strategy, we propose AdvSumm for mitigating bias in text summarization. As shown in Figure 2, there are three major components in AdvSumm. First is an encoder E, which maps the input text x into a continuous latent space providing a text embedding $E(x) = e_x$. The second component, Perturber, makes perturbations to e_x in the direction that maximally increases the loss function $\mathcal{L}$, thereby targeting regions of the embedding space that are most likely to degrade the model's generation quality. The continuous text representation e_x allows us to generate an adversarial sample e'_x using the gradient-based methods such as PGD (Madry et al., 2019). The last component decoder D maps back the perturbed e'_x back to the input space. We use the Transformer (Vaswani et al., 2023) encoder and decoder architectures and optimize the cross-entropy loss $\mathcal{L}(y, D(e'_x))$, where y is the ground truth bias-free summary. This process is outlined in Algorithm 1.

We use the Fast Gradient Signed Method (FGSM) (Goodfellow et al., 2015) to build the Perturber component, which is a cheaper single-step variant of PGD, with the number of iterations $t = 1$. The perturbed embedding is generated with the following embedding update in FGSM:

$$e'_x = e_x + \epsilon \cdot sgn\left(\frac{\partial \mathcal{L}}{\partial e_x}\right) \tag{1}$$

where, $sgn(.)$ represents the sign of the quantity and ϵ captures the attack strength.

Embedding e_x and model's predicted output $\hat{y} = D(e_x)$ are generated with a forward pass of the Encoder and Decoder respectively. The model's predicted output $\hat{y}$ along with the ground-truth summary y are used to compute the loss $\mathcal{L}(y, \hat{y})$. The sign of the gradient of the computed loss function $\mathcal{L}$ is then used to modify e_x to e'_x using equation 1. Similar to the Stochastic Gradient Descent parameter optimization technique, equation 1 identifies the steepest ascent of loss as a function of embedding e_x. Therefore, the Perturber modifies the embedding such that the update direction results in the largest increase in the generation loss. Since this change leads to the highest increase in loss, this adversarial embedding e'_x must lead to the worst generated summary among all the embeddings in the ϵ ball radius of e_x. In the training procedure, this perturbed embedding e'_x is then used to jointly train the Encoder and Decoder.

Dataset	Type	#Train	#Test	#Val
XSUM	News Summ	203,577	11,305	11,301
Wiki-Nationality	Nationality Hallucination	0	71,763	0
Multi-Neus	Multi-polar News Summ	2,453	307	307

Table 1: Statistics of datasets used in this work.

4 Experiments

4.1 Datasets

We evaluate AdvSumm on two existing bias summarization benchmarks: name-nationality bias (Ladhak et al., 2023) and political framing bias (Lee et al., 2022). These datasets allow us to assess the generalization capability of our method across different kinds of bias. Specifically, name-nationality bias primarily arises from hallucinated tokens—where the model incorrectly introduces demographic attributes (e.g., inferring nationality based on names)—while political framing bias involves more subtle language choices at the document level, reflecting ideological leanings. By addressing both token-level and discourse-level biases, we demonstrate the broader applicability of our approach. Dataset statistics are summarized in Table 1.

Name-Nationality Bias. For assessing name-nationality hallucination, we use the Wiki-Nationality dataset (Ladhak et al., 2023) which was constructed by altering entity names in articles to associate them with different nationalities, without changing other biographical details. This was done to assess whether models will use an incorrect/assumed nationality in the summary just based on the person's name.

Framing Bias. We explore framing bias with the Neutral multi-news Summarization (NeuS) dataset (Lee et al., 2022), which comprises triplets of left, right, and center-slanted news articles paired with neutral summaries focused on the facts in the articles.

4.2 Metrics

Name-Nationality Bias. We calculate the hallucination rate as the proportion of articles where the model incorrectly attributes a nationality in the generated summary which is different from the nationality in the input document. The aim of our approach is to reduce the hallucination rate and hence, reduce the spurious association of names to specific nationalities.

Attack Strength	ROUGE-1	Ar_+	Ar_-	Ar_{sum}
0	44.81	2.19	1.07	3.26
10^{-3}	44.38	1.82	0.93	2.75
10^{-2}	41.07	0.55	0.29	0.84
10^{-1}	14.52	0.29	0.16	0.45

Table 2: Flan-T5 on Multi-Neus with different degrees of attack strength.

Framing Bias. Following Lee et al. (2022), we use arousal scores from the Valence-Arousal-Dominance (VAD) lexicon (Mohammad, 2018), which provides valence (v), arousal (a), and dominance (d) annotations for a list of words. The positive arousal score (Ar_+) and negative arousal score (Ar_-) are defined as the summed arousal values of words with positive and negative valence, respectively, based on the VAD annotations. The combined arousal score (Ar_{sum}) is the sum of Ar_+ and Ar_-. The goal of AdvSumm is to mitigate political framing in generated summaries by minimizing both Ar_+ and Ar_-, while preserving overall summarization quality.

Summarization Quality. We utilize ROUGE (Lin, 2004) scores to measure the summarization quality. We report ROUGE-1 in our results.

4.3 Settings

Models. We use two encoder-decoder transformer models for the text summarization task: BART-large (Lewis et al., 2019) and Flan-T5 base (Chung et al., 2022). BART is a denoising autoencoder pre-trained with a corrupted text reconstruction objective, making it well-suited for generation tasks. In contrast, Flan-T5 builds on the T5 architecture (Raffel et al., 2023) and is further instruction-tuned on a broad mixture of tasks, enabling better generalization to unseen instructions and objectives. This contrast allows us to evaluate the robustness and generalization capabilities of AdvSumm across models with different pre-training strategies. We leave it to future work to adapt our Perturber component to decoder-only LLM architectures.

For name-nationality bias, models are fine-

Model	ROUGE-1 ↑	American↓	Asian↓	African↓	European↓	Overall↓
BART	43.45	0.84	13.41	0.92	7.55	5.61
Flan-T5	39.99	0.03	2.39	0.06	0.57	0.76
Back-Trans (BART)	41.91	1.08	8.69	1.32	4.75	3.96
AdvSumm (BART)	40.02	0.40	4.42	0.27	1.96	1.76
AdvSumm (Flan-T5)	37.86	0.02	2.33	0.16	0.38	0.72

Table 3: Hallucination rate over multiple countries in Wiki-Nationality dataset. Am represents American, Af African, As Asian and Ovr is the Hallucination rate over all countries. AdvSum improves the Hallucination rate while maintaining similar ROUGE-1 scores.

tuned on the XSUM news summarization dataset (Narayan et al., 2018) and evaluated on the Wiki-Nationality benchmark. For framing bias, we adopt the fine-tuning scheme of Lee et al. (2022) for fine-tuning on the training split of the Multi-Neus dataset. AdvSumm applies adversarial training with the perturber component during this fine-tuning stage.

Baselines. We compare AdvSumm against two baselines: (i) models fine-tuned on the same data without the perturber component, and (ii) a data augmentation using back-translation. For the latter, training data is augmented by paraphrasing input texts via back-translation from German, effectively doubling the training set size while keeping the targets unchanged (Cao and Wang, 2021). We evaluate the effectiveness of this back-translation-based generalization strategy against our adversarial generalization method (AdvSumm).

Implementation. We experiment on an NVIDIA A100 GPU with 40 GB VRAM. We finetune all models using a learning rate of 5e-5 with AdamW optimizer and 10% warm-up steps. Maximum input length is set to 1024 for XSUM and 512 for Multi-Neus. The maximum output generation length is taken as 142 along with a beam size of 6 for Wiki-Nationality and a generation length of 250 with a beam size of 4 is used for Multi-Neus. Generation configurations (like input, output lengths, beam sizes, etc.) are adopted directly from Lee et al. (2022). Other hyperparameters like the number of epochs are tuned on validation splits. All results are reported on the test split. For the baselines, we use English-to-German and German-to-English translation models provided by Fairseq for back-translation.

We tune the attack strength of the Perturber Component by varying the value of ϵ in equation 1. A higher value of ϵ gives the Perturber higher freedom to change the embedding e_x but also leads to a greater change in text semantics, which will lead to a drop in summarization performance. For practical implications, ϵ behaves like a "knob" for controlling the amount of bias while trading off the summarization quality. We show the tuning results of Flan-T5 on the Multi-Neus dataset in Table 2 with the ROUGE-1 and Arousal scores. We observe a drop in bias as well as summarization quality as we increase the value of attack strength. We find $\epsilon = 0.01$ to be optimal, which we use for further experiments.

5 Results

We present our empirical findings on the two types of biases in this section.

5.1 Name-Nationality Bias

The results on the Name-Nationality benchmark are illustrated in Table 3, which shows a comparative analysis between the baseline models and our proposed approach, AdvSumm, focusing on region-specific hallucination rates as discussed in Ladhak et al. (2023), as well as ROUGE-1 scores on the XSum evaluation sets to compare summarization quality.

In the Name-Nationality setting, AdvSumm significantly lowers hallucination rates in summaries across American, Asian, and European contexts compared to base models. It effectively reduces overall hallucination rates underscoring the effectiveness of AdvSumm in enhancing the fidelity of summarization models, ensuring more reliable summaries across diverse geopolitical landscapes while maintaining competitive ROUGE-1 scores.

AdvSumm's lower ROUGE-1 scores on the test sets, as shown in Table 3, align with prior research findings that Adversarial Training, while enhancing model robustness, can reduce performance on clean data (Madry et al., 2019). This tradeoff is expected in data augmentation techniques, where the goal is to improve model resilience (reduce bias) while

Models/ Settings	Framing Bias Metrics			Salient Info
	$Ar_{+} \downarrow$	$Ar_{-} \downarrow$	$Ar_{sum} \downarrow$	ROUGE-1 $\uparrow$
BART	1.33	0.76	2.09	45.94
Flan-T5	2.19	1.07	3.26	44.81
Back-Trans	1.40	0.77	2.17	**46.51**
AdvSum(BART)	0.59	0.33	0.92	43.01
AdvSum(Flan-T5)	**0.55**	**0.29**	**0.84**	41.07

Table 4: Results of AdvSum on of Multi-Neus dataset. An attack strength of 0.01 is used for AdvSum. BART-large is used for training on Back-Translated data. Ar stands for Arousal.

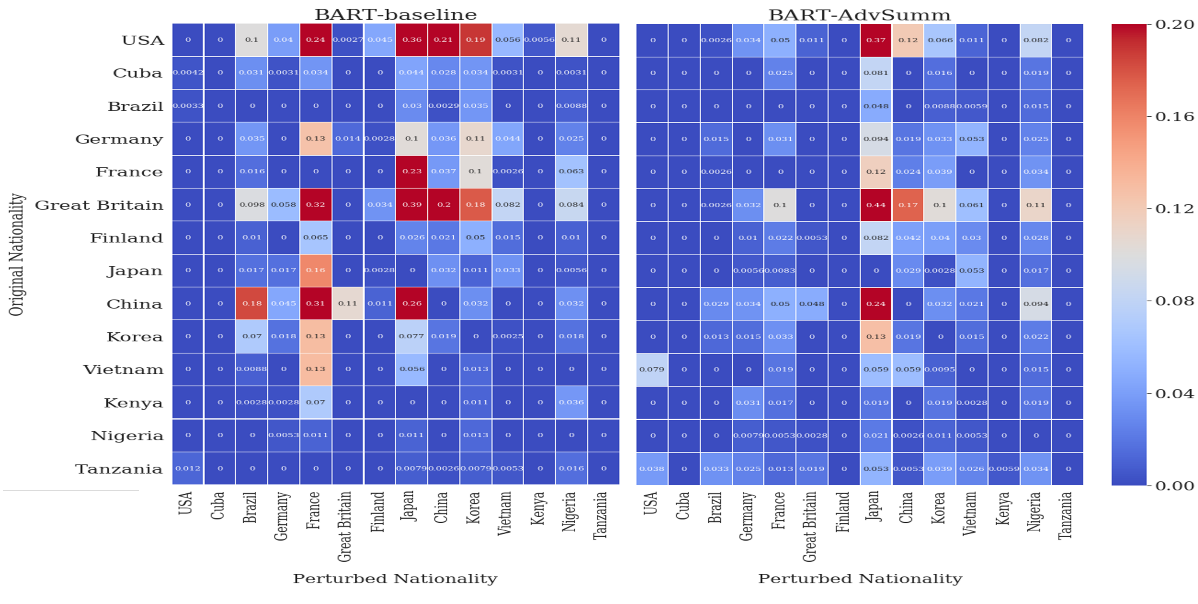

Figure 3: Hallucination rate for BART baseline and one trained using AdvSum. Red corresponds to higher, and Blue corresponds to lower hallucination rate.

minimizing the performance drop on clean datasets.

5.2 Framing Bias

Evaluations on the Multi-Neus benchmarks are outlined in Table 4. We report the Framing Bias Metric consisting of positive Arousal Score A_{+}, negative Arousal Score A_{-}, and their sum A_{sum}. We also report ROUGE-1 for capturing the summarization quality by each setting. We also show the bias evaluations of the back-translation-based data-augmentation approach in Table 4.

On the Multi-Neus dataset, we see the least biased summaries in the case of AdvSumm on Flan-T5, with the lowest positive and negative Arousal scores. Both generalization approaches (ours and back-translation) outperform Flan-T5, which supports the hypothesis on bias mitigation with robustness. AdvSumm surpasses the back-translation approach by 1.4 absolute points on Ar_{sum}, while taking a slight dip in ROUGE-1 scores. We also note that our end-to-end adversarial training approach is more computationally efficient than back-translation, given the time taken by dual-translation and the double training steps of summarization fine-tuning.

We also observe consistently lower bias scores across both benchmarks when using the Flan-T5 architecture. Flan-T5 benefits from instruction tuning on a diverse set of tasks, including ethical reasoning and instruction following. We hypothesize that this additional tuning phase not only enhances zero-shot generalization but also better aligns the model with human expectations, helping it avoid spurious biases inherited from pre-training.

5.3 Error Analysis

For name-nationality bias, we report a heatmap as shown in Figure 3 where the hallucination rate for all combinations of countries is calculated. In alignment with the numerical results, hallucination rates for Asian countries as perturbed nationalities are significantly higher for the Bart baseline than our approach AdvSumm. We notice that, however, for a few combinations like Great Britain-Japan,

AdvSum	Texas Church Shooting: A gunman opened fire at a church in Texas on Sunday, killing two people and wounding three others.
Source News	Shooting at Texas Church Leaves 2 Parishioners Dead, Officials Say: A gunman opened fire at a church in Texas on Sunday morning, killing two people with a shotgun before a member of the church's volunteer security team fatally shot him, the authorities said. About 250 people were inside the auditorium of the West Freeway Church of Christ in White Settlement, near Fort Worth, when the gunman began shooting just before communion, said Jack Cummings, a minister at the church. Mr. Cummings said the gunman was "acting suspiciously" before the shooting and drew the attention of the church's security team.

Table 5: An example of positive arousal generated news. AdvSum hallucinates the text in red color. Each of the three examples contains <Title>:<Article>. The Center news article is shown in Source News.

Generated News	Trump to End DACA: President Trump will announce on Tuesday that he is ending a controversial program that protects nearly 800,000 young undocumented immigrants from deportation, media reports indicated late Sunday.
Neutral News	Reports Say DACA Is Over: President Trump will announce on Tuesday that he is ending a controversial program that protects nearly 800,000 young undocumented immigrants from deportation, media reports indicated late Sunday.

Table 6: A generated news summary compared to neutral news. Each example contains <Title>:<Article>

Vietnam-USA, there is a slight increase in the hallucination rate.

For framing bias, most of the biased generation is still a result of model hallucination. The example shown in Table 5 shows the text in red color, which is hallucinated by the model. The "wounding of three others" is not mentioned in the source article. Additionally, current Framing bias metrics fail to capture the context around lexicons. An example is shown in Table 6, where the positive arousal score given by the Lexicon-based metric is zero, which is clearly wrong, looking at the title of the generated news.

6 Conclusions

In this work, we introduced AdvSumm, a domain-agnostic adversarial training framework for bias mitigation in text summarization. Motivated by the limitations of existing bias mitigation strategies—particularly their domain-specific nature and difficulty generalizing across different types of biases—we reformulated bias reduction as a generalization problem, tackled through adversarial robustness. By introducing the Perturber module to apply embedding-level adversarial perturbations during fine-tuning, we demonstrated that AdvSumm effectively reduces both token-level biases (e.g., name-nationality associations) and document-level biases (e.g., political framing) without compromising summarization quality. Empirical results on benchmark datasets highlight that AdvSumm outperforms standard transformers and back-translation baselines, offering a unified and scalable solution for fairer text generation.

Limitations

Our study focuses on bias mitigation in text summarization using encoder-decoder transformer architectures. However, many recent summarization systems adopt decoder-only architectures, where directly applying the Perturber component in its current form is not straightforward. Future work could explore extending adversarial perturbations to individual layers of the transformer decoder, enabling the approach to generalize to decoder-only models as well.

Ethics Statement

We conduct our evaluations using publicly available datasets that do not contain personally sensitive information or toxic content. One important ethical consideration is that developing robust summarization systems, as proposed in this paper, contributes to ongoing efforts to reduce harmful biases in natural language generation systems by mitigating biases inherited from pre-training data. For example, prior work has shown that biased news framing can contribute to political polarization (Han and Federico, 2017), and name-nationality associations can reinforce harmful stereotypes in text generation (Ladhak et al., 2023).

By improving the robustness of summarization models, our approach takes a step toward addressing these issues. However, we acknowledge that our work does not evaluate all forms of bias that may arise in text summarization tasks, nor does it fully evaluate potential side effects of the approach, such as its impact on other aspects of faithfulness or other types of bias in summarization. Future research should explore these broader impacts to ensure that summarization systems are both fair and faithful across different contexts and biases.

Acknowledgments

One of the authors is supported by the National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia University Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Another author is supported by Amazon and Columbia's Center of Artificial Intelligence Technology (CAIT) PhD student fellowship.

References

Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. 2009. *Robust optimization*, volume 28. Princeton university press.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, pages 43–57. Springer.

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized data pre-processing for discrimination prevention.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. 2020. Empirical risk minimization under fairness constraints.

Ragnhild Eg, Özlem Demirkol Tønnesen, and Merete Kolberg Tennfjord. 2023. A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, 9:100253.

Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Leveraging annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Jiyoung Han and Christopher M. Federico. 2017. Conflict-framed news, self-categorization, and partisan polarization. *Mass Communication and Society*, 20(4):455–480.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Maximilian Kaufmann, Yiren Zhao, Ilia Shumailov, Robert Mullins, and Nicolas Papernot. 2022. Efficient adversarial training with data pruning.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. 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.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. 2018. Domain generalization via conditional invariant representation.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

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.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks.

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.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. 2021. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 466–477, New York, NY, USA. Association for Computing Machinery.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2024. Bias in news summarization: Measures, pitfalls and corpora.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Martin Wessel, Tomás Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB - the first media bias identification benchmark task and dataset collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Eric Wong and J. Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope.

Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Ziqiang Cao, Sujian Li, and Hua Wu. 2022a. Frsum: Towards faithful abstractive summarization via enhancing factual robustness. *arXiv preprint arXiv:2211.00294*.

Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Sujian Li, and Yajuan Lyu. 2022b. Precisely the point: Adversarial augmentations for faithful and informative text generation. *arXiv preprint arXiv:2210.12367*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Yue Xing, Qifan Song, and Guang Cheng. 2021. On the generalization properties of adversarial training. In *International Conference on Artificial Intelligence and Statistics*, pages 505–513. PMLR.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. 2021. Improved ood generalization via adversarial training and pretraining.

Huan Zhang, Shiqi Wang, Kaidi Xu, Yihan Wang, Suman Jana, Cho-Jui Hsieh, and Zico Kolter. 2022. A branch and bound framework for stronger adversarial attacks of ReLU networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26591–26604. PMLR.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Adam Štorek, Mukur Gupta, Noopur Bhatt, Aditya Gupta, Janie Kim, Prashast Srivastava, and Suman Jana. 2025. Xoxo: Stealthy cross-origin context poisoning attacks against ai coding assistants.