What Constitutes a Faithful Summary? Preserving Author Perspectives in News Summarization

Yuhan Liu[*♣] **Shangbin Feng**[*♠] **Xiaochuang Han**[♠]
Vidhisha Balachandran[♡] **Chan Young Park**[♡] **Sachin Kumar**[◇] **Yulia Tsvetkov**[♠]
[♣]Xi'an Jiaotong University, [♠]University of Washington
[♡]Carnegie Mellon University, [◇]Allen Institute for AI
lyh6560@stu.xjtu.edu.cn; shangbin@cs.washington.edu

Abstract

In this work, we take a first step towards designing summarization systems that are faithful to the author's opinions and perspectives. Focusing on a case study of *preserving political perspectives in news summarization*, we find that existing approaches alter the political opinions and stances of news articles in more than 50% of summaries, misrepresenting the intent and perspectives of the news authors. We thus propose P^{3}SUM, a diffusion model-based summarization approach controlled by political perspective classifiers. In P^{3}SUM, the political leaning of a generated summary is iteratively evaluated at each decoding step, and any drift from the article's original stance incurs a loss back-propagated to the embedding layers, steering the political stance of the summary at inference time. Extensive experiments on three news summarization datasets demonstrate that P^{3}SUM outperforms state-of-the-art summarization systems and large language models by up to 11.4% in terms of the success rate of stance preservation, with on-par performance on standard summarization utility metrics. These findings highlight the lacunae that even for state-of-the-art models it is still challenging to preserve author perspectives in news summarization, while P^{3}SUM presents an important first step towards evaluating and developing summarization systems that are faithful to author intent and perspectives.[1]

1 Introduction

What constitutes a faithful summary? In addition to preserving factual consistency—the focus of much prior work (Kryscinski et al., 2020; Goyal and Durrett, 2020; Wang et al., 2020a; Pagnoni et al., 2021; Feng et al., 2023a; Tam et al., 2023)—a good summarization system should preserve the *writer's voice*—the style, intent, and points of view conveyed by the authors. However, such subtle pragmatic cues are harder to extract and control for by existing models (Borji, 2023), and it remains underexplored whether existing summarization systems generate summaries that are *faithful* to the opinions and perspectives of the authors. What's worse, language models (LMs) inevitably contain political biases and such biases could further impact downstream tasks (Kumar et al., 2022a; Feng et al., 2023b). Specifically in the task of summarization, instead of "de-biasing" and generating only neutral summaries, we argue that a good summarization system should *preserve the political perspectives* of the news authors in generated summaries for opinion diversity and democratic discourse.

To this end, we first evaluate to what extent summarization systems and LLMs preserve political stances in generated summaries, by employing a state-of-the-art political perspective evaluator (Liu et al., 2022d) to quantify the gap between news articles and summaries. (§2) We identify that existing summarization systems and LLMs *do* alter opinions and perspectives in the original document, resulting in shifting stances in more than 50% of summaries, with around 25% drifting to the partisan extremes (Figure 1). This highlights a new, underexplored concern with current LLMs as they fail to preserve the intents and perspectives of the authors of news documents during summarization, potentially misinforming the readers.

To address this issue, we propose P^{3}SUM, a summarization model aiming to **P**reserve the **P**olitical **P**erspectives of news articles. (§3) P^{3}SUM employs a non-autoregressive diffusion language model with modular control capabilities to steer the generated summary towards the same perspective of the news article. Specifically, we first fine-tune a diffusion language model (Mahabadi et al., 2023; Han et al., 2023b,a) on summarization data. During inference, the generated summary is evaluated by a

[*]These authors contributed equally to this work.
[1]Our code and models will be released at https://github.com/lyh6560new/P3Sum.

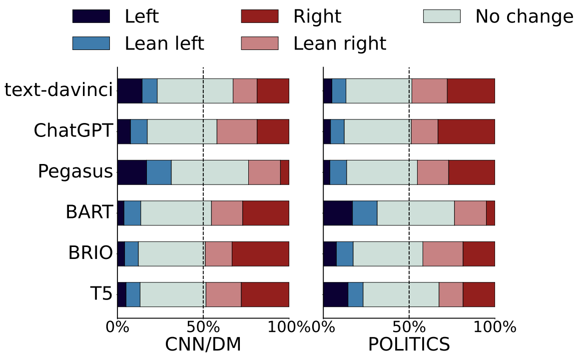

Figure 1: Changes in political stances between the summary and the article. The political perspective classifier produces *left*, *center*, or *right* labels for each text sequence. Left (or Right) indicates a shift in summary stance towards left (or right) by 2 units while Lean Left (Or Lean Right) indicates a shift by 1 unit. No change indicates that there is no difference in the political leaning of the summary and the context. **Our study shows that existing approaches alter the stances of news articles in more than 50% of cases across both datasets**.

political stance classifier (Liu et al., 2022d) at each step, compared to the target stance in the source document while summary generation is steered towards the target stance. Our primary motivation to use diffusion models is that they allow us to apply the stance classifier on the whole summary at each decoding step, rather than on a prefix generated autoregressively (Kumar et al., 2022b). As an inference-time approach based on diffusion models and controllable text generation (Kumar et al., 2021; Li et al., 2022a; Han et al., 2023a,b; Mahabadi et al., 2023; Austin et al., 2021; Strudel et al., 2022; Dieleman et al., 2022), P^{3}SUM alleviates the need for additional training or pretraining, handles news articles from different ideological stances, and is seamlessly compatible with future state-of-the-art classifiers on political perspective detection.

Extensive experiments on three news datasets demonstrate that P^{3}SUM greatly outperforms baselines in preserving the political stances of news articles while maintaining good summarization utility. Specifically, P^{3}SUM is at least 11.4%, 6.2%, and 2.7% better in perspective preservation on CNN/DM (Nallapati et al., 2016), XSUM (Narayan et al., 2018), and POLITICS (Liu et al., 2022d), outperforming popular summarization systems (Raffel et al., 2020; Liu et al., 2022b; Zhang et al., 2020) and large language models (Touvron et al., 2023; Penedo et al., 2023; Chiang et al., 2023). In addition, P^{3}SUM obtains ROUGE scores and abstractiveness metrics that are only slightly lower than state-of-the-art systems, while qualitative analysis highlights P^{3}SUM's effectiveness in generating high-quality, perspective-preserving summaries. We envision P^{3}SUM as a first step towards summarization systems that are faithful to the intents and perspectives of the authors.

2 Examining Perspective Preservation

Given a news article, the generated summary should preserve the authors' political perspectives in the document. However, existing models and systems are not designed to control for author intent or perspectives, and we first investigate to which extent summarization systems and large language models alter author perspectives in the generated summaries.

To this end, we measure the political leaning of the generated summaries and compare them to the political stances of original articles, using 500 randomly chosen news articles from the CNN/DM (Nallapati et al., 2016) and POLITICS (Liu et al., 2022d) datasets. We use a political perspective evaluator (Liu et al., 2022d) to quantify political stances of summaries and news articles (mapping text sequences to *left*, *center*, or *right*), investigating the change in political leanings with six summarization models and LLMs: GPT-3.5 (TEXT-DAVINCI-003), CHATGPT (GPT-3.5-TURBO), PEGASUS (Zhang et al., 2020), BART (Lewis et al., 2020), BRIO (Liu et al., 2022b), and T5 (Raffel et al., 2020). We then determine the political perspective gap between the summary and the news article.

As shown in Figure 1, current summarization systems struggle to provide faithful summaries and significantly alter political perspectives. Concretely, the political stance of the generated summary is different from the news article in more than 50% of cases across different models, while around 25% drift to partisan extremes. As a result, how to develop summarization approaches that are faithful to the authors' perspectives in the news document remains an open research question.

3 P^{3}SUM

We propose P^{3}SUM, a diffusion model that can steer the political stance of the generation towards the news article during inference time with an off-the-shelf classifier. Given a news article $\boldsymbol{d}$, P^{3}SUM aims to generate a summary $\boldsymbol{s}$ that preserves the original political stance of the article. We first finetune a diffusion-based language model on summarization datasets. At decoding time, we employ a political stance classifier to steer the generated summary by incorporating the gradient information of the classifier, ensuring that the political stance of the generated summary is consistent with the original article.

3.1 Diffusion Model Finetuning

At a high level, a diffusion model performs forward diffusion by adding noise to the original data and then learns to reconstruct the input. During inference time, we use the learned model to iteratively reconstruct from noisy representations and obtain high-quality generations. To preserve the political stance, we modify the decoding process by incorporating the gradients from an external political classifier iteratively to guide the model generation.

Continuous data representation Following Han et al. (2023a), we define a function $\text{logits-initialization}(\cdot)$ to obtain a logits representation over the model's vocabulary $\mathcal{V}$, mapping each discrete tokens of the news context and summary into continuous space. We map a token w to $\tilde{\boldsymbol{w}} \in \{-K, +K\}^{|V|}$ as follows:

$$\tilde{w}^{(j)} = \begin{cases} +K \text{ when } w = V^{(j)} \\ -K \text{ when } w \neq V^{(j)} \end{cases}$$

where $V^{(j)}$ denotes the j-th token in the vocabulary and K is a pre-defined scalar hyperparameter.

Forward diffusion For each passage $\boldsymbol{d}$ and gold summary $\boldsymbol{s}$, we concatenate them to form a sequence $\boldsymbol{w} = (w_1, \ldots, w_L)$. We adopt non-autoregressive modeling (Mahabadi et al., 2023) which feeds the entire sequence into the model to better handle long article contexts. Let $\mathbf{S}_0 = (\tilde{\boldsymbol{w}}_1, \ldots, \tilde{\boldsymbol{w}}_L) \in \{\pm K\}^{L\times|V|}$ be the logit representations of $\boldsymbol{w}$. Each step in the forward diffusion derives $\mathbf{S}_t$ by: $\mathbf{S}_t = \sqrt{\bar{\alpha}_t}\mathbf{S}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$ where $t \in (1, T)$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, K^2\mathbf{I})$, and $\bar{\alpha}_t \to 0$ as $t \to T$ following a predefined schedule. At step T, $\text{sm}(\mathbf{S}_T)$ are fully noisy simplexes over V (we use sm as a shorthand for softmax).

Reverse process Based on the noisy representation $\mathbf{S}_t$ (or noisy simplex $\text{sm}(\mathbf{S}_t)$) and a current timestep t, we learn to reverse the forward process by predicting the original representation $\mathbf{S}_0$ with our model $\text{Transformer}_\theta$. The predicted outputs are the output logits from the Transformer model θ, denoted as $\hat{\mathbf{S}}_\theta(\mathbf{S}_t, t)$.

$$\hat{\mathbf{S}}_\theta(\mathbf{S}_t, t) = \text{Transformer}_\theta(\text{sm}(\mathbf{S}_t), t) \quad (1)$$

We also apply self-conditioning (Chen et al., 2022) with a 50% probability during prediction, recomputing $\mathbf{S}_t$ in Eq. 1 by:[2]

$$\mathbf{S}_t = \frac{1}{2}(\mathbf{S}_t + \hat{\mathbf{S}}_\theta(\mathbf{S}_t, t))$$

Loss function After obtaining the model prediction $\hat{\mathbf{S}}_\theta(\mathbf{S}_t, t)$, we employ a cross-entropy loss between this predicted representation of $\mathbf{S}_0$ and the target summary tokens $\boldsymbol{w}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{t,\mathbf{S}_0}\left[-\sum_{i\in\mathbf{s}} \log p_{\boldsymbol{\theta}}(w_i|\mathbf{S}_t, t)\right] \\ &= \mathbb{E}_{t,\mathbf{S}_0}\left[-\sum_{i\in\mathbf{s}} \log \text{sm}[\hat{\mathbf{S}}_\theta(\mathbf{S}_t, t)]_{w_i}\right] \end{aligned}$$

where $\log p_{\boldsymbol{\theta}}(\cdot|\cdot)$ denotes the cross-entropy loss over the output logits of the transformer model θ that we are learning,[3] and $i \in \mathrm{s}$ denotes whether this token belongs to summary s.

3.2 Perspective-Guided Decoding

A diffusion language model generates the output sequence non-autoregressively by initializing a noise sequence $\mathbf{S}_T$ and iteratively refining it through $\mathbf{S}_{t+1}, \mathbf{S}_t, \ldots, \mathbf{S}_0$.

Given an article as input, we initialize the summary as a noisy sequence $\mathbf{S}_T$ where each token is represented as a logit sampled from the normal distribution $\mathcal{N}(\mathbf{0}, K^2\mathbf{I})$. Using our learned model $\boldsymbol{\theta}$, We first obtain an estimated output reconstructing from $\mathbf{S}_T$:

$$\hat{\mathbf{S}}_{\text{sc},T} = \hat{\mathbf{S}}_{\boldsymbol{\theta}}(\mathbf{S}_T, T), \quad (2)$$

Self-Conditioning Mahabadi et al. (2023) observe that self-conditioning (Chen et al., 2022) can improve the consistency between the model predictions and given context. Following their setting, for all steps $t < T$, we perform self-conditioning by mixing and leveraging the predictions from the previous time step in the current step. Let $\mathbf{S}_{t+1}$ denotes the incoming logits at t from the previous time step $t+1$, and $\hat{\mathbf{S}}_{sc,t+1}$ denotes the original

[2]Please refer to Mahabadi et al. (2023) for details.

[3]For more details, please refer to Han et al. (2023a) and Han et al. (2023b).

estimation of the logits at time step $t+1$. We perform self-conditioning by computing the average of these representations and then pass to the model $\boldsymbol{\theta}$ for a prediction:

$$\hat{\mathbf{S}}_{\text{sc},t} = \hat{\mathbf{S}}_{\boldsymbol{\theta}}(\frac{\mathbf{S}_{t+1} + \hat{\mathbf{S}}_{\text{sc},t+1}}{2}, t+1)$$

Modular control We employ political bias classifiers to steer the generated summary toward the stances of the news article. To guide P^{3}SUM to generate summaries with a target political leaning $y \in \{$*left*, *center*, *right*$\}$, we use an external stance classifier $f_\phi(\cdot)$ that maps texts to the three stance labels and update our previous prediction $\hat{\mathbf{S}}_{\text{sc},t}$ at each timestep t guided by the gradients from the political stance classifier.

$$\hat{\mathbf{S}}_{\text{ctr},t} = \hat{\mathbf{S}}_{\text{sc},t} + \lambda \nabla_{\hat{\mathbf{S}}_{\text{sc},t}} f_\phi(y \mid \text{sm}(\hat{\mathbf{S}}_{\text{sc},t}))$$

where λ is a hyperparameter governing the intensity of stance steering and the parameters of ϕ are frozen. This enables P^{3}SUM to iteratively steer the political stances of the generated summary toward the news article. P^{3}SUM employs a modular *plug and control* paradigm so that any off-the-shelf political bias classifier[4] could be seamlessly integrated.

Logits projection To obtain the almost one-hot logits similar to the initial data distribution, we further project logits $\hat{\mathbf{S}}_{\text{ctr},t}$ at the end of every iteration following (Han et al., 2023b):

$$\hat{\mathbf{S}}^{(j)}_{\text{proj},t} = \begin{cases} +K \text{ if } j\text{=top-}p\text{-sampling}(\hat{\mathbf{S}}_{\text{ctr},t}) \\ -K \text{ otherwise} \end{cases}$$

where top-p is the hyperparameter for nucleus sampling (Holtzman et al., 2019). After projecting $\hat{\mathbf{S}}_{\text{ctr},t}$ to $\hat{\mathbf{S}}_{\text{proj},t}$, we add a noise according to the forward diffusion schedule and pass the representation $\mathbf{S}_t$ as the incoming logits for the next iteration $t-1$:

$$\mathbf{S}_t = \sqrt{\bar{\alpha}_t}\hat{\mathbf{S}}_{\text{proj},t} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$$

So the decoding process can be summarized as iteratively denoising logits $\mathbf{S}_T$ to obtain $\mathbf{S}_{t+1}, \mathbf{S}_t, \ldots, \mathbf{S}_0$, and $\mathbf{S}_0$ is the final summary. At time step t, we first mix the noisy logits $\mathbf{S}_{t+1}$ and the model estimation $\hat{\mathbf{S}}_{sc,t+1}$ from time step $t+1$ (self-conditioning) and obtain a model estimation for step t: $\hat{\mathbf{S}}_{sc,t}$. Then, we apply the classifier to predict the perspective for the current estimation $\hat{\mathbf{S}}_{sc,t}$ and compare it with a target stance y. The difference between the prediction and the target stance is backpropagated to steer the logits $\hat{\mathbf{S}}_{\text{ctr},t}$. After that, we project the logits $\hat{\mathbf{S}}_{\text{ctr},t}$ to $\hat{\mathbf{S}}_{\text{proj},t}$ and add Gaussian noise to derive $\mathbf{S}_t$. Such process is repeated T times with $\mathbf{S}_0$ as the final representation. The final summary is obtained by converting $\operatorname{argmax} \mathbf{S}_0$ to natural language tokens.

$$\begin{aligned}
\hat{\mathbf{S}}_{\text{sc},t} &= \hat{\mathbf{S}}_{\boldsymbol{\theta}}(\frac{\mathbf{S}_{t+1} + \hat{\mathbf{S}}_{\text{sc},t+1}}{2}, t+1) \\
\hat{\mathbf{S}}_{\text{ctr},t} &= \hat{\mathbf{S}}_{\text{sc},t} + \lambda \nabla_{\hat{\mathbf{S}}_{\text{sc},t}} f_\phi(y \mid \text{sm}(\hat{\mathbf{S}}_{\text{sc},t})) \\
\hat{\mathbf{S}}_{\text{proj},t} &= \text{logits-projection}(\hat{\mathbf{S}}_{\text{ctr},t}) \\
\mathbf{S}_t &= \sqrt{\bar{\alpha}_t}\hat{\mathbf{S}}_{\text{proj},t} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t
\end{aligned}$$

4 Experiments

4.1 Experimental Settings

Datasets We adopt three news datasets: CNN/DM (Nallapati et al., 2016), XSUM (Narayan et al., 2018), and POLITICS (Liu et al., 2022d). Since there are no ground truth summaries provided in POLITICS, we employ the GPT-3.5-TURBO model from OpenAI API to generate reference summaries similar to Zhang et al. (2023).

Baselines We compare P^{3}SUM with two types of baselines: 1) *summarization systems*, specifically BRIO (Liu et al., 2022b), PEGASUS (Zhang et al., 2020), and T5 (Raffel et al., 2020). 2) *large language models*, specifically Vicuna (Chiang et al., 2023), Falcon (Penedo et al., 2023), and Llama-2 (Touvron et al., 2023).[5] For each baseline, we employ two modes: *without preservation*, where the baseline is directly used for summarization; *with preservation*, where we prepend instructions to encourage stance preservation.[6]

Implementation We employ the encoder-only ROBERTA-BASE (Liu et al., 2019) as the backbone of P^{3}SUM's diffusion component. To preserve perspectives at inference time, we leverage the political bias classifier from POLITICS (Liu et al., 2022d), which measures the political stance of the generation and compares it with the original stance at each decoding step. This allows a loss term measuring the political stance difference to backpropagate to the embedding layers, penalizing per-

[4]We assume a common tokenizer.

[5]We test them in the zero-shot setting.

[6]For similar baselines of controllable text generation such as Liu et al. (2021a), we do not compare them with our method since the classifier we use is a discriminator, not a generator as required by the paper.

Method	Pres.	Model Size	POLITICS		CNN/DM		XSUM	
			Suc↑	Dist↓	Suc↑	Dist↓	Suc↑	Dist↓
T5	✗	200M	44.10	0.35	47.13	0.38	50.53	0.35
BRIO	✗	400M	44.95	0.35	48.65	0.37	29.19	0.49
PEGASUS	✗	568M	44.19	0.36	44.03	0.37	25.40	0.51
VICUNA	✗	7B	51.90	0.30	39.83	0.35	45.35	0.32
FALCON	✗	40B	45.34	0.35	39.20	0.38	31.90	0.41
LLAMA2	✗	70B	46.93	0.34	39.26	0.39	43.03	0.35
T5	✓	200M	47.29	0.34	41.83	0.40	47.97	0.38
BRIO	✓	400M	42.15	0.38	46.98	0.38	30.96	0.48
PEGASUS	✓	568M	42.38	0.36	43.78	0.38	31.28	0.48
VICUNA	✓	7B	52.16	0.29	43.49	0.34	36.02	0.36
FALCON	✓	40B	44.28	0.34	41.72	0.35	38.60	0.36
LLAMA2	✓	70B	45.08	0.34	39.51	0.38	51.54	**0.30**
P^{3}SUM (ours)	✓	125M	**53.57**	**0.26**	**54.18**	**0.31**	**54.75**	0.33

Table 1: Performance of political perspective preservation on the three datasets. "Pres." indicates whether the model is instructed to preserve stances or not. ↑ and ↓ indicate whether the metric should be high or low. P^{3}SUM outperforms all baseline approaches that are 1.6x to 560x larger on five of the six settings across the three datasets.

spective inconsistencies. We provide full details of P^{3}SUM training and inference in Appendix A.

Evaluation We define two metrics to evaluate the success of preserving political stances in the summary using the political stance classifier that maps text sequences to a bias label $f_{bias}(\cdot)$: str $\rightarrow$ $\{-1, 0, 1\}$ representing left, center, and right-leaning. 1) *Success Rate* (Suc): $\frac{1}{|\mathcal{D}|}\sum_{\boldsymbol{d}\in\mathcal{D}} \mathbb{1}(f_{bias}(\boldsymbol{d}) = f_{bias}(\boldsymbol{s}))$, where $\mathbb{1}(\cdot)$ denotes the indicator function and $\mathcal{D}$ denotes the full dataset. 2) *Stance Distance* (Dist): $\frac{1}{|\mathcal{D}|}\sum_{\boldsymbol{d}\in\mathcal{D}} |f_{bias}(\boldsymbol{d}) - f_{bias}(\boldsymbol{s})|$. While Suc examines whether the stance of the summary is consistent with the article, Dist further evaluates how far the perspective of summaries drifts from the news documents. For summarization utility evaluation, we employ Rouge-1/2/L scores (Lin, 2004) and summarization abstractiveness scores (Chan et al., 2021).

4.2 Results

Preserving Author Perpsectives Table 1 demonstrates that P^{3}SUM achieves the highest average success rate as well as the lowest stance distance across five of the six settings, outperforming baselines that are 1.6x to 560x larger. For success rate, we surpass the second-best method by 2.7%, 11.4%, and 6.2% respectively on the POLITICS,CNN/DM, and XSUM datasets. This suggests that the combination of diffusion language models and plug-in political bias classifiers offers a promising approach to preserving political perspectives in news summarization.

For large language model baselines that perform text summarization in a zero-shot setting, we observe that adding instructions for stance preservation produces mixed effects on their performance. For example, the instructions work for FALCON on CNN/DM but are counterproductive on POLITICS. We hypothesize that large language models struggle to grasp the concept of preserving political opinions off-the-shelf, potentially influenced by their internal notion of political leanings that is often biased and inaccurate (Shaikh et al., 2022; Feng et al., 2023b). However, with an explicit classifier-based gradient steering paradigm, P^{3}SUM successfully advances the ability to preserve political perspectives in generated summaries.

Summarization Utility We evaluate P^{3}SUM and baselines on CNN/DM and POLITICS by comparing them to reference summaries and present results in Tables 2 and 3. Table 2 demonstrates that P^{3}SUM achieves Rouge scores that are on-par with state-of-the-art approaches, while Table 3 shows that P^{3}SUM is producing abstractive and concise summaries. Together these results demonstrate that P^{3}SUM gets better at preserving political opinions and perspectives without greatly sacrificing summarization quality.

Qualitative Analysis In Table 4, we present an example news article from the POLITICS dataset, where models produce summaries with different po-

Method	POLITICS				CNN/DM			
	R1	R2	R-L	R-avg	R1	R2	R-L	R-avg
T5	38.31	18.04	27.82	33.07	40.82	18.30	28.64	29.25
BRIO	47.91	24.24	33.12	35.09	46.21	22.04	31.36	33.20
PEGASUS	40.62	19.36	29.64	29.87	42.70	19.69	29.76	30.72
VICUNA	21.33	8.84	14.78	14.98	13.20	3.48	8.51	8.40
FALCON	18.77	4.32	11.28	11.46	15.59	3.17	9.43	9.40
LLAMA2	30.93	12.98	20.72	21.54	22.21	6.75	13.89	14.28
P^3SUM (ours)	37.48	16.50	26.01	26.66	41.12	18.20	27.73	29.02

Table 2: Rouge scores on POLITICS and CNN/DM. Though the decoding process is steered by classifier gradients to preserve political stances, P^3SUM's summarization utility is still competitive among baselines.

Method	POLITICS	CNN/DM	XSUM
T5	9.02	8.61	7.15
BRIO	5.17	4.11	3.16
PEGASUS	6.76	3.80	6.46
VICUNA	3.98	2.64	1.50
FALCON	1.77	0.83	0.65
LLAMA2	3.99	2.20	1.29
P^3SUM (ours)	6.32	2.59	2.93

Table 3: Abstractiveness scores (Chan et al., 2021), the lower the better. P^3SUM successfully produces concise summaries that are competitive with existing approaches while improving perspective preservation.

litical leanings. The original article takes a mostly neutral stance, analyzing the electorate and voter issues. However, T5 generates a strongly left-leaning summary by priming the hostility from Republicans and focusing on incorrect facts such as a Republican Senate to support its argument.[7] BRIO instead makes a right-leaning pitch by highlighting the challenges looming for the incoming administration. In contrast, P^3SUM maintains a neutral standpoint, summarizing the demographic differences in the 2020 election and preserving the original article's political stance, as confirmed by the stance classifier.

5 Related Work

Text Summarization and Factuality Evaluation Research on neural text summarization has produced models and systems that are capable of generating fluent and informative summaries (Liu and Lapata, 2019; Balachandran et al., 2021; Rothe et al., 2021; Narayan et al., 2021; Bhattacharjee et al., 2023; Chen et al., 2023b; He et al., 2023; Liu et al., 2023b; Chen et al., 2023a), given documents from various domains such as news articles (Fabbri et al., 2019; Liu et al., 2022a; Bahrainian et al., 2022), scientific literature (Goldsack et al., 2022), social media and dialogue (Tang et al., 2022; Liu et al., 2022c). However, it remains challenging to generate summaries that are factually consistent with the given document (Cao et al., 2018; Balachandran et al., 2022), resulting in the research area of factuality evaluation. Existing works propose benchmarks to evaluate the factuality of generated summaries (Pagnoni et al., 2021; Tang et al., 2023), develop factuality evaluation models and metrics (Wang et al., 2020b; Kryscinski et al., 2020; Nan et al., 2021; Goyal and Durrett, 2021; Ribeiro et al., 2022; Utama et al., 2022; Laban et al., 2022; Feng et al., 2023a; Luo et al., 2023), and improve the factuality of generated summaries (Aharoni et al., 2023; Liu et al., 2023a). Recent studies suggest that state-of-the-art large language models (Goyal et al., 2022; Bhaskar et al., 2022) are capable of achieving remarkable factuality in text summarization. However, while LLMs are capable of generating summaries that are factually faithful, our work demonstrates that they struggle to generate summaries that are faithful to the authors' original opinions and perspectives (Figure 1). As a result, we propose P^3SUM, an important first step towards summarization systems that preserve the authors' intents, opinions, and perspectives in the generated summary.

Understanding the Social and Political Biases of Language Models Extensive research has demonstrated that machine learning models could encode and exhibit social and political biases (Zhao et al., 2018; Blodgett et al., 2020; Bender et al.,

[7] In 2020, Democrats narrowly won control of the senate with a tie-breaking vote from the Vice President.

Context	Model	Summary	Stance
Biden ... will confront a divided country beset by an unprecedented and complex set of difficulties ... Election returns and exit polls revealed sharp differences between men and women and white and minority Americans.... His response to these challenges will be limited by a Republican Senate, a solidly conservative Supreme Court majority, hostility from Trump supporters ... Biden enjoyed a big edge with non-white Americans while white voters stuck with the incumbent... (**center**)	Ours	Election returns and exit polls reveal sharp differences between men and women and white....	center ✓
	T5	Biden ... will be limited by a Republican Senate, a solidly conservative Supreme Court majority, hostility from Trump supporters. ...	left ✗
	BRIO	... Biden must confront the pandemic, rebuild the economy and address climate change ...	right ✗

Table 4: A qualitative example of generated summaries from different approaches. Existing summarization systems often alter the political perspective by presenting partial facts or making up non-existing statements. Our method successfully preserves the original perspective by presenting only the main idea and facts in the original article.

2021; Goldfarb-Tarrant et al., 2021; Jin et al., 2021; Shaikh et al., 2022; Li et al., 2022b). Existing works mainly analyze biases expressed in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Kurita et al., 2019), token probabilities (Borkan et al., 2019; Bordia and Bowman, 2019; Liu et al., 2021b), model performance discrepancy (Hardt et al., 2016; Feng et al., 2023b), and generated texts (Kumar et al., 2022a). Specifically for political biases, several studies have been proposed to probe LLMs (Bang et al., 2021; Liu et al., 2021b; Feng et al., 2023b), evaluate the political leaning of texts (Li and Goldwasser, 2019; Feng et al., 2021; Zhang et al., 2022; Liu et al., 2022d; Qiu et al., 2022), and pretraining LMs on partisan corpora (Jiang et al., 2022). Annotator (Geva et al., 2019; Sap et al., 2019; Davani et al., 2022; Sap et al., 2022; Gordon et al., 2022; Chen and Zhang, 2023) and data bias (Park et al., 2018; Dixon et al., 2018; Dodge et al., 2021; Harris et al., 2022) are commonly attributed as the cause of LM biases, while existing works also established that LM biases could propagate into downstream tasks and cause fairness issues (Gonen and Webster, 2020; Li et al., 2020; Dodge et al., 2021; Feng et al., 2023b; Steed et al., 2022). In this work, we uniquely focus on the task of news summarization: while existing LM-based summarization approaches generate summaries that are not consistent with the political stances of the article, we propose P^{3}SUM to steer the perspective of the summary through iterative controllable text generation.

Controllable text generation In text summarization, controllable text generation can generate summaries with given entities, predefined lengths, and more (Chan et al., 2021; He et al., 2020; Li et al., 2022a). More generally, inference-time methods can be used to steer the generation process by altering the output probability distribution at decoding time (Dathathri et al., 2019; Krause et al., 2021; Yang and Klein, 2021; Liu et al., 2021a; Lu et al., 2021; Pascual et al., 2021; Kumar et al., 2021; Qin et al., 2022; Kumar et al., 2022b; Mireshghallah et al., 2022). Particularly, Han et al. (2023a) and Mahabadi et al. (2023) leverage diffusion-based methods that applies inference-time control through off-the-shelf classifiers. In this work, we further explore the summarization setup using diffusion models that preserve political leanings in the decoding process.

6 Conclusion

We demonstrate that existing summarization systems and LLMs struggle to preserve the authors' political perspectives in news summarization. We present P^{3}SUM, a diffusion-based summarization model that improves political perspective preservation by iteratively guiding the decoding process with an external political stance classifier. Extensive experiments demonstrate that P^{3}SUM outperforms large language models and summarization systems in producing summaries faithful to the political stances of news documents while maintaining competitive summarization utility.

Limitations

Time Overhead While our proposed P^{3}SUM is better than existing summarization systems and LLMs at preserving authors' political perspectives in the generated summaries, it comes at the cost of inference time subject to the classifier control component at the decoding time of diffusion models. We employ 1000 decoding steps to refine a generated summary so that it is consistent with the news articles' perspectives and stances, which adds to inference-time computational costs.

Political Bias Classifier We employ POLITICS (Liu et al., 2022d), an LM-based political bias classifier to iteratively steer the political stances of the generated summary. While it successfully helps to preserve author perspectives, it only provides coarse-grained categorical political leanings (left/-center/right) and is not 100% accurate. However, we argue that our proposed P^{3}SUM is seamlessly compatible with future political bias classifiers that are more accurate and fine-grained.

Ethics Statement

Although P^{3}SUM's intended use case is to preserve author perspectives in news summarization, there is a potential risk for misuse: malicious attempts could be made with P^{3}SUM to steer the political leaning of the generated summary towards the hyperpartisan extremes, furthering societal divides and deepening polarization. Therefore, we plan to establish access permission to the fine-tuned P^{3}SUM weights to ensure that it is only used for research purposes.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

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

Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime G Carbonell, and Yulia Tsvetkov. 2021. Structsum: Summarization via structured representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2575–2585.

Yejin Bang, Nayeon Lee, Etsuko Ishii, Andrea Madotto, and Pascale Fung. 2021. Assessing political prudence of open-domain chatbots. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 548–555.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Adithya Bhaskar, Alexander R Fabbri, and Greg Durrett. 2022. Zero-shot opinion summarization with gpt-3. *arXiv preprint arXiv:2211.15914*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. Crosssum: Beyond english-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained Markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232.

Quan Ze Chen and Amy X Zhang. 2023. Judgment sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement. *arXiv preprint arXiv:2305.01615*.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. 2022. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*.

Xiuying Chen, Guodong Long, Chongyang Tao, Mingzhe Li, Xin Gao, Chengqi Zhang, and Xiangliang Zhang. 2023a. Improving the robustness of summarization systems with dual augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6846–6857, Toronto, Canada. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2023b. Unisumm and summzoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. 2022. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023a. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281*.

Shangbin Feng, Zilong Chen, Wenqian Zhang, Qingyao Li, Qinghua Zheng, Xiaojun Chang, and Minnan Luo. 2021. Kgap: Knowledge graph augmented political perspective detection in news media. *arXiv preprint arXiv:2108.03861*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023b. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature.

In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, 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.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023a. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11575–11596, Toronto, Canada. Association for Computational Linguistics.

Xiaochuang Han, Sachin Kumar, Yulia Tsvetkov, and Marjan Ghazvininejad. 2023b. Ssd-2: Scaling and inference-time fusion of diffusion language models. *arXiv preprint arXiv:2305.14771*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 789–798.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Pengcheng He, Baolin Peng, Song Wang, Yang Liu, Ruochen Xu, Hany Hassan, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. 2023. Z-code++: A pre-trained language model optimized for abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5095–5112, Toronto, Canada. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. CommunityLM: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Proc. Findings of EMNLP*.

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.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022a. Language generation models can cause harm: So what can we do about it? an actionable survey. *arXiv preprint arXiv:2210.07700*.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34:14542–14554.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022b. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

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

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.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022a. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Anton Ragni, Shi Wang, and Jie Fu. 2022b. HERB: Measuring hierarchical regional bias in pre-trained language models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 334–346, Online only. Association for Computational Linguistics.

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.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021a. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proc. ACL*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021b. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740.

Yang Liu, Chenguang Zhu, and Michael Zeng. 2022a. End-to-end segmentation-based news summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 544–554, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023a. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022c. Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 703–710.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022d. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 4288–4299, Online. Association for Computational Linguistics.

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

Rabeeh Karimi Mahabadi, Jaesung Tae, Hamish Ivison, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. 2023. Tess: Text-to-text self-conditioned simplex diffusion. *arXiv preprint arXiv:2305.08379*.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proc. ACL*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

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

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. *arXiv preprint arXiv:1808.08745*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

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.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *ArXiv*, abs/2202.11705.

Changyuan Qiu, Winston Wu, Xinliang Frederick Zhang, and Lu Wang. 2022. Late fusion with triplet margin objective for multimodal ideology prediction and analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9720–9736, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 140–145.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

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.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, et al. 2022. Self-conditioned embedding diffusion for text generation. *arXiv preprint arXiv:2211.04236*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Prasetya Utama, Joshua Bambrick, Nafise Sadat Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level nli examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776.

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

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

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. In *Proc. NAACL*.

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

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Wenqian Zhang, Shangbin Feng, Zilong Chen, Zhenyu Lei, Jundong Li, and Minnan Luo. 2022. KCD: Knowledge walks and textual cues enhanced political perspective detection in news media. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4140, Seattle, United States. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20.

Pres.	Prompt
✗	You are an abstractive summarizer that follows the output pattern. Please write a summary for the document. Document:
✓	You are an abstractive summarizer that follows the output pattern. You should always respect the political bias in the original article.Please write a summary for the document. Document:

Table 5: Prompts for large language models, where "Pres." is the abbreviation for whether it belongs to the preservation setting.

Method	POLITICS	CNN/DM
TEXT-DAVINCI	74.93	84.44
CHATGPT	96.15	89.35

Table 6: Factuality score for LLM-generated summaries.

A Experiment Details

During fine-tuning on summarization, we use a leaning rate of $3e-5$, We fine-tuned for 20000 steps.

For decoding, we use top- $p = 0.95$ suggested in Han et al. (2023a) and 1000 diffusion steps according to Mahabadi et al. (2023).

We implement P^{3}SUM on a server using Tesla V100 GPU with 32 GB memory, 16 CPU cores, and 377GB memory for the experiments.

The backbone of our model is ROBERTA-BASE. It's noticeable that both P^{3}SUM and the model in (Liu et al., 2022d) use ROBERTA-BASE, and thus they share the same tokenizer. Therefore, as mentioned in (Han et al., 2023a), they can be used for control in an off-the-shelf manner.

For POLITICS, there are no human-written summaries. Therefore, we take the summarization of GPT-TURBO as the ground truth. The details are in the appendix E

With CNN/DM as a popular dataset in text summarization, we aim to test how well P^{3}SUM can perform traditional summarization tasks. However, not all the news articles in the CNN/DM are within the political discipline, which is inappropriate for political leaning preservation. Therefore, we leverage the POLITICS dataset(Liu et al., 2022d), which consists of political news with labels of political leaning.

B Political Biases of Gold Summaries

We measure the political perspective of gold summary and demonstrate its change from the given

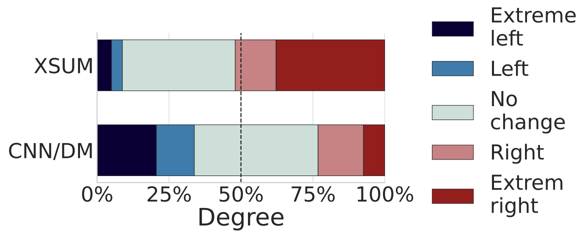

Figure 2: Changes in political stances between the gold summary and the article.

article on CNN/DM and XSUM. Although these human-written or selected summaries are considered as gold standards for summarization tasks and used to train the model, they can hardly preserve the political perspectives themselves. And the summarization systems, fine-tuned on these data, can be affected by the political bias.

C Understanding Political Instructions in Prompts

The prompt we use for zero-shot inference for large language models are listed in the table 5.

D Qualitative Analysis (cont.)

Although P^{3}SUM achieves the highest performance on the datasets, it can also fail in certain cases. We present one failure in table 7.

E Selecting Criteria

Because there aren't gold summaries in the POLITICS(Liu et al., 2022d) dataset, we use model-generated summaries for calculating rouge scores. We prompt the TEXT-DAVINCI and CHATGPT, and compare factuality and overall rouge scores.

We calculate the factuality score of summaries by Feng et al. (2023a) and present the scores in table 6. It is shown that CHATGPT has a higher level of faithfulness.

Choosing TEXT-DAVINCI and CHATGPT as reference summaries respectively, we calculate the rouge scores respectively on POLITICS dataset and present the results in table 8.

We can see that most models achieve higher rouge scores when selecting CHATGPT to generate gold summaries, which implies a higher agreement.

Context	Model	Summary	Stance
For months, Republican leaders have been uniform in their insistence that they would allow everyone's taxes to rise if the rich did not get to keep their Bush-era tax breaks. Mr. Obama has proposed continuing the tax cut for the 98 percent of taxpaying families . . . Republicans have demanded tax cuts for all, and, so far, not a single Republican leader has lined up behind Mr. Boehner's concession. Ultimately, the case for the top-level tax cuts is increasingly shaky. If Republicans are the least bit serious about reducing the deficit, they have to acknowledge that doing so requires additional revenues. . . (**left**)	Ours	Republican leaders have been ready to maintain Bush-era tax breaks to continue tax rates. Mr. Obama, who has earned less than $250,000, will keep up with extra revenue at top rates. . . .	right ✗
	T5	The case for the top-level tax cuts is increasingly shaky. If Republicans are the least bit serious about reducing the deficit, they have to acknowledge that doing so requires additional revenues.	left ✓
	BRIO	. . . Republicans have demanded tax cuts for all, . . . If Republicans are serious about reducing the deficit, they have to acknowledge that doing so requires additional revenues.	left ✓

Table 7: A qualitative example of news summarization.

Method	text-davinci as gold				ChatGPT as gold			
	R-1	R-2	R-L	R-avg	R-1	R-2	R-L	R-avg
T5	28.40	11.20	21.66	20.42	36.35	17.50	27.62	27.16
BRIO	31.11	13.66	23.25	22.67	47.91	24.24	33.12	35.09
PEGASUS	26.10	9.40	19.37	18.29	40.62	19.36	29.64	29.87

Table 8: Comparison of rouge scores using TEXT-DAVINCI or CHATGPT as gold summaries.