Mitigating Adversarial Perturbations in Abstractive Text Summarization through Bias-Amplified Task Vector Negation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) for abstractive text summarization exhibit strong lead bias, disproportionately relying on initial sentences, making them susceptible to adversarial perturbations that exploit this positional dependence. Recent work has shown that this lead bias can be exploited through perturbation-based sentence exclusion attacks, where minor character-, word-, sentence-, or document-level modifications cause models to omit important information from generated summaries. In this work, we introduce ***Bias-Amplified Task Vector Negation*** **(BA-TVN)**, a parameter-space defense for mitigating positional shortcut reliance in summarization models. BA-TVN isolates excessive lead-position reliance through bias-amplified fine-tuning and suppresses the resulting positional shortcut direction via scaled vector negation. We evaluate BA-TVN across multiple summarization architectures and datasets. Results show consistent reductions in positional reliance and improved robustness against sentence exclusion attacks, with substantially larger Lead Bias Ratio (LBR) reductions than prior bias-mitigation, unlearning, and input-level robustness baselines. BA-TVN also preserves summarization quality, with minimal changes in lexical, semantic, and factual quality metrics. These findings suggest that bias-amplified parameter-space editing offers a practical and architecture-agnostic mechanism for mitigating positional shortcut behaviors in summarization models.

1 Introduction

Text summarization plays a critical role in condensing extensive documents into concise representations that directly influence downstream decision-making. Recent advances in large-scale pre-trained models, including GPT-5 (OpenAI, 2025), Claude (Anthropic, 2025), PaLM (Anil et al., 2023), etc., have enabled summarization systems to operate effectively across various domains, such as news aggregation (Huang et al., 2024), policy analysis (Chen et al., 2025), and scientific reporting (Cachola et al., 2020). Despite strong summarization performance, recent work has shown that these models exhibit strong positional dependencies, often over-relying on lead sentences when generating summaries (Nallapati et al., 2017; Thota and Nilizadeh, 2024). In particular, small character-, word-, sentence-, or document-level perturbations applied to early sentences can induce sentence exclusion behavior, causing models to omit important information from generated summaries.

Existing robustness approaches in NLP primarily focus on input-level defenses such as adversarial training (Miyato et al., 2017), adversarial detection (Shen et al.), and text purification (Moraffah et al., 2024), which are designed mainly for classification settings and transfer poorly to summarization. More fundamentally, these approaches treat adversarial perturbations as input-level artifacts to detect or remove, rather than addressing the positional dependencies encoded in model parameters. Machine unlearning (Bourtoule et al., 2021) offers a complementary perspective, enabling targeted removal of undesirable behaviors from model parameters without full retraining.

In this work, we introduce ***Bias-Amplified Task Vector Negation*** **(BA-TVN)**, a parameter-space defense to mitigate reliance on positional shortcuts in abstractive summarization. Unlike prior applications of task vectors that target semantic or social biases, such as gender or sentiment associations, positional shortcut reliance is a structural generation behavior arising from how summarization models allocate content selection across document positions. This makes it difficult to isolate from standard summarization data, where lead-position information is often naturally correlated with summary relevance. To address this challenge, BA-TVN constructs a bias-amplified task vector by deliberately strengthening lead reliance during

controlled fine-tuning, allowing the corresponding shortcut behavior to be captured in parameter space. We then suppress this direction through scaled vector negation, producing a model with reduced positional dependence while preserving general summarization quality. Rather than treating adversarial perturbations as input-level artifacts to sanitize at inference time, our approach edits the model behavior that enables perturbation-based sentence exclusion attacks.

We evaluate our defense across eleven summarization models and six datasets spanning news, encyclopedic, and scientific domains. The models include both encoder-decoder architectures, such as BART-Large, Pegasus-Large, etc., and decoder-only instruction-tuned models, such as LLaMA-3-8B, Qwen-3-4B-Instruct, Gemma-3-it, etc. Results show consistent reductions in positional reliance and improved robustness against perturbation-based sentence exclusion attacks while preserving summarization quality. Compared to prior bias-mitigation (Grenander et al., 2019; Gupta et al., 2025; Lee et al., 2022), unlearning baselines (Bourtoule et al., 2021; Izzo et al., 2021), and input-level text purification-based defenses (Li et al., 2023; Pruthi et al., 2019; Makepeace and contributors, 2021; Si et al., 2021), our approach achieves larger reductions in Lead Bias Ratio (LBR) with minimal impact on summarization quality. We further study the effects of task-vector scaling and bias amplification strategies through ablation analysis. We also evaluate robustness under adaptive perturbations applied to non-lead positions, finding that TVN-defended models recover summarization quality close to clean baselines even under such strategies.

The primary contributions of our work are as follows: (1) We study how positional shortcut reliance in summarization models contributes to perturbation-induced sentence exclusion, where small input modifications cause models to omit important source content from generated summaries. (2) We introduce Bias-Amplified Task Vector Negation (BA-TVN), a parameter-space defense that constructs bias-amplified summarization signals to isolate and suppress excessive lead-position reliance. (3) We evaluate BA-TVN across six datasets and eleven model architectures, demonstrating consistent reductions in positional reliance, improved robustness to sentence exclusion attacks, and minimal impact on lexical, semantic, and factual quality metrics. (4) We compare BA-TVN against prior bias-mitigation, generic unlearning, and input-level robustness baselines, showing stronger and more stable reductions in Lead Bias Ratio (LBR). (5) We release our full implementation and evaluation pipeline, URL: https://anonymous.4open.science/r/Defense_Adversarial_Perturbations_Summarization.

2 Related Work

Mitigating Adversarial Perturbations. Existing defenses primarily rely on adversarial training (Goodfellow et al., 2014; Madry et al., 2018; Miyato et al., 2017), detection-based filtering (Rosenfeld et al., 1996), or text purification (Li et al., 2023), and are designed mainly for classification settings. These approaches treat perturbations as input-level artifacts rather than addressing the underlying positional dependencies that contribute to perturbation sensitivity. They also transfer poorly to summarization, which requires preserving factual content, coherence, and content selection across the full document.

Machine Unlearning for Model Editing. Machine unlearning methods (Cao and Yang, 2015; Bourtoule et al., 2021) remove learned behaviors from model parameters without full retraining. Early work focused on exact unlearning via data partitioning (Bourtoule et al., 2021), while later methods use gradient-based updates (Thudi et al., 2022) or second-order approximations (Golatkar et al., 2020) to approximate data removal at lower cost. These methods typically remove the influence of specific training examples or learned associations after training. In contrast, BA-TVN uses unlearning-inspired parameter editing as a proactive defense for abstractive summarization. Rather than removing specific examples, it constructs bias-amplified summarization signals to expose and suppress positional shortcut reliance directly in parameter space. To our knowledge, BA-TVN is first task-vector-based defense for mitigating perturbation-induced sentence exclusion caused by positional shortcut reliance in abstractive summarization.

Bias Mitigation in Summarization. Position bias, where models over-rely on information from specific input locations, has been observed across NLP tasks (Kedzie et al., 2018; Ko et al., 2020) and manifests as lead bias in summarization (Nallapati et al., 2017). Prior mitigation approaches use training-time interventions such as auxiliary

objectives or shuffled pretraining (Kryściński et al., 2019), focusing mainly on summary coverage rather than perturbation sensitivity. Recent task-vector methods mitigate semantic or social biases, such as gender or sentiment associations, by constructing vectors from corpora expressing stereotyped or harmful associations (Dige et al., 2024). In contrast, BA-TVN targets perturbation-induced sentence exclusion caused by positional shortcut reliance, a structural content-selection behavior tied to document position. As described in Section 4.1, we therefore construct bias-amplified summarization signals to make this positional shortcut extractable as a task vector.

3 Threat Model

Figure 1 illustrates this threat model and defense setting. We consider an attacker with black-box access to a deployed summarization model who can modify input documents at inference time. The attacker's goal is to induce sentence exclusion attacks by exploiting the model's inherent lead bias, causing it to omit important content, particularly from early sentences, in the generated summary. The attacker applies adversarial perturbations to lead sentences at multiple granularities, including character-level (e.g., insertions, deletions, homoglyph substitutions), word-level (e.g., synonym replacement or deletion), sentence-level (e.g., paraphrasing or reordering), and document-level manipulations (e.g., sentence or paragraph reordering) (Thota and Nilizadeh, 2024). An example of such attacks is demonstrated in Appendix 10.1. ***Defender Goals and Capabilities.*** The defender's goal is to reduce the model's excessive reliance on lead-position shortcuts so that adversarial perturbations are less likely to cause sentence exclusion, while preserving normal summarization quality. Defenders are model developers with full parameter access during development, who can apply our defense before deployment, requiring no inference-time overhead or input preprocessing. This is in contrast to reactive approaches such as adversarial training or input sanitization.

4 Methodology

Text summarization models exhibit a well-documented tendency to over-rely on initial sentences when generating summaries, a phenomenon known as lead bias (Nallapati et al., 2017). While this behavior is beneficial in settings where important information naturally appears early (Grenander et al., 2019; Steen and Markert, 2024), it also creates positional shortcut reliance that makes models sensitive to perturbations in lead content (Thota and Nilizadeh, 2024). Our goal is to reduce this excessive positional reliance directly in parameter space while preserving valid use of lead information.

4.1 Task Vector Negation for Positional Shortcut Mitigation

Task vectors are based on the observation that specific model behaviors can be represented as directions in parameter space. Prior task-vector work has used biased text corpora to construct vectors for semantic or social biases, such as stereotypical associations involving protected groups (Dige et al., 2024). In those settings, the target behavior is expressed through lexical or semantic associations in the fine-tuning data. Our setting is different: positional shortcut reliance is not tied to a specific demographic attribute or semantic association, but to where summary-relevant information appears in the document. Because lead sentences are often genuinely important in summarization, standard datasets do not cleanly separate valid lead usage from excessive positional dependence. We therefore introduce a bias-amplified task-vector construction procedure. Instead of using naturally occurring biased sentences, we construct summarization examples in which summary-relevant content is deliberately concentrated in the first few sentences. Fine-tuning on this dataset induces exaggerated lead reliance, and the resulting parameter difference captures the positional shortcut direction. Negating this direction reduces excessive positional dependence while preserving general summarization behavior.

Figure 2 illustrates our defense methodology through four main components. Given an original model f_{θ_0} and a model f_{θ_t} fine-tuned to express a target behavior, the task vector is defined as the parameter difference $\tau_t = \theta_t - \theta_0$. Adding this vector can reinforce the target behavior, while subtracting a scaled version can suppress it:

$$\theta_{\text{edited}} = \theta_0 - \lambda \tau_t.$$

We first construct a dataset $\mathcal{D}_{\text{bias}}$ that intentionally concentrates summary-relevant content in the first few sentences. Fine-tuning the original model on this dataset produces a biased model $f_{\theta_{\text{bias}}}$ with exaggerated lead reliance. We then compute the

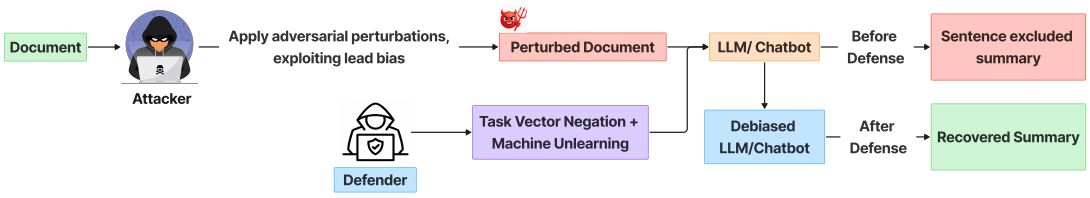

Figure 1: Threat model illustrating inference-time adversarial perturbations and development-time defense

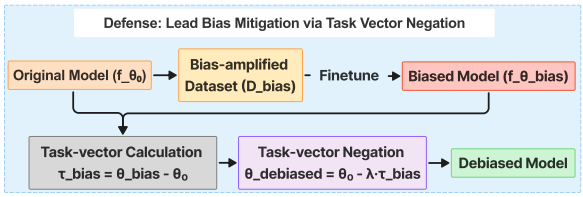

Figure 2: Methodology for BA-TVN

positional shortcut vector:

$$\tau_{\text{pos}} = \theta_{\text{bias}} - \theta_0.$$

Finally, we suppress this direction through scaled negation:

$$\theta_{\text{TVN}} = \theta_0 - \lambda \tau_{\text{pos}}.$$

Thus, the key contribution is not the vector arithmetic alone, but the construction of a summarization-specific bias-amplified signal that makes positional shortcut reliance identifiable as a parameter-space direction. The scaling coefficient λ allows control over the bias–quality tradeoff.

4.2 Bias Amplification

A key challenge in applying task vectors to lead-bias mitigation is that naturally occurring lead bias is not strong enough for reliable task-vector extraction. Prior work shows that standard summarization datasets often exhibit approximately 60–70% lead bias due to journalistic writing conventions (See et al., 2017; Zhu et al., 2021; Kryściński et al., 2019; Nallapati et al., 2017). However, our ablations show that this natural signal is insufficient: reliable extraction requires bias concentration approaching 90–95%. We therefore construct bias-amplified training data to create a stronger directional signal.

We implement two complementary amplification strategies. The first, **document restructuring**, modifies source documents while keeping reference summaries unchanged. Summary-relevant sentences are identified using ROUGE overlap and semantic similarity, then reordered to appear in the first 2–3 positions. Early sentences are further augmented with positional markers (`[LEAD]`, `[KEY]`) and trained with position-weighted loss functions that assign exponentially higher weight to lead content. The second, **summary modification**, keeps source documents unchanged and instead modifies reference summaries to explicitly incorporate content from the first 2–3 sentences, using extractive techniques to identify and integrate lead information. Both strategies are supplemented with synthetic document-summary pairs achieving the required 90–95% bias concentration. Document restructuring tends to produce stronger positional signals for heterogeneous document structures, while summary modification is computationally more efficient. Either approach alone is sufficient for effective task vector extraction; we use document restructuring for all primary experiments.

4.3 QLoRA Adaptation

For decoder-only models, full fine-tuning is computationally expensive. We therefore apply QLoRA (Dettmers et al., 2023) during bias-amplified fine-tuning, keeping base weights frozen in 4-bit precision and training only low-rank adapter weights. The task vector is computed from the LoRA adapter updates, which capture the positional bias direction induced during fine-tuning. Scaled negation is then applied to these adapter weights to suppress lead-biased behavior while preserving general summarization capability.

5 Experimental Setup

5.1 Models and Datasets

Models: We evaluate across eleven summarization models covering encoder-decoder and decoder-only architectures. Encoder-decoder models include BART-Large (Lewis et al., 2019), T5-Small (Raffel et al., 2020), Pegasus-Large (Zhang et al., 2020), and FLAN-T5-Large (Chung et al., 2024). Decoder-only models include LLaMA-3-8B (Dubey et al., 2024), Qwen-2.5-3B-Instruct (Qwen et al., 2024), Qwen-3-4B-Instruct (Yang et al., 2025), Mistral-7B (Jiang et al., 2023), and Vicuna-13B (Chiang et al., 2023). We

additionally evaluate on gemma-3-it (Kamath et al., 2025) and gpt-oss-20b (Agarwal et al., 2025) to assess generalization to more recent models. As a result, we focus on open-source models and do not consider proprietary APIs such as GPT-4 (Hurst et al., 2024) or Claude (Anthropic, 2025), where internal parameters are inaccessible. Encoder-decoder models are fine-tuned in full precision, and decoder-only models use QLoRA for computational efficiency. **Datasets:** We conduct experiments on six diverse datasets spanning multiple domains. For news summarization, we use CNN/ DailyMail (Chen et al., 2016) for single-document news summarization containing nearly 300K news articles, and MultiNews (Fabbri et al., 2019) containing 44,972 training examples of multi-document news articles. For encyclopedic content, we use WikiSum (Cohen et al., 2021) with structured informational articles, containing over 1M article-summary pairs. For scientific literature, we evaluate on ArXiv (nearly 215k) and PubMed (nearly 133k) (Cohan et al., 2018) for computer science and biomedical papers respectively, and Multi-XScience (Lu et al., 2020) comprising 30,369 training examples from multiple scientific domains. This diversity ensures our defenses are robust across varying writing styles, document lengths, and domain-specific vocabularies.

5.2 Evaluation Protocol

Task vectors are extracted using bias-amplified fine-tuning on 10,000 examples per dataset, which we verify is sufficient for stable extraction across the 1k–10k range. Robustness and quality are evaluated on 10,000 held-out test instances per dataset that are never used during task-vector extraction. All bias-inducing fine-tuning runs use the same hyperparameters: learning rate 2×10^{-5}, batch size 8, and gradient accumulation steps 2, and are executed on NVIDIA A6000 GPUs. For decoder-only models, QLoRA reduces the additional fine-tuning cost to approximately 1–3 GPU-hours per model, and edited models incur no inference-time preprocessing or runtime overhead. TVN is largely insensitive to hyperparameters beyond the scaling coefficient λ, provided that bias amplification produces a stable task vector.

5.3 Baselines

We compare our proposed TVN defense against two categories of baselines that represent alternative approaches to mitigating undesirable model behavior. **Bias Mitigation Baselines.** These include multi-stage training with auxiliary losses (Grenander et al., 2019), adversarial bias mitigation via AdvSumm (Gupta et al., 2025), and task-level neutral summarization via NeuS (Lee et al., 2022). These methods use training-time objective modifications and do not explicitly isolate positional dependencies in parameter space. **Unlearning-Based Baselines.** Unlike TVN, these baselines reduce learned behavior through retraining or iterative optimization. For a fair comparison, we adapt them to the same bias-amplified signals used by TVN: (i) *exact unlearning* via SISA-style isolation (Bourtoule et al., 2021), which retrains on partitions excluding bias-amplified examples; and (ii) *approximate unlearning* via gradient-based updates (Izzo et al., 2021), which reduces the influence of bias-amplified training signals. **Adversarial Robustness Baselines.** We also compare against input-level robustness defenses, including text adversarial purification (Li et al., 2023), scRNN (Pruthi et al., 2019), LanguageTool normalization (Makepeace and contributors, 2021), and adversarial data augmentation (Si et al., 2021). For adversarial data augmentation, models are fine-tuned on clean and perturbed examples generated using the perturbation strategies of (Thota and Nilizadeh, 2024).

5.4 Metrics and Adaptive-Attack Strategies

Lead Bias Ratio (LBR).

We quantify positional shortcut reliance using Lead Bias Ratio (LBR), defined as:

$$\text{LBR} = \frac{\text{ROUGE-1}_{F1}(\text{First 3 sentences, Summary})}{\text{ROUGE-1}_{F1}(\text{Entire document, Summary})}. \quad (1)$$

LBR measures how much of the generated summary overlaps with lead content relative to the full source document. Higher values indicate stronger reliance on lead sentences, while lower values indicate more balanced content selection across document positions. We use the same ROUGE-1 F1 tokenization and stemming settings in both numerator and denominator. To ensure that LBR is not tied to the fixed three-sentence window, we also compute variants using percentage-based lead windows covering the first 10–20% of each document.

Adversarial Robustness Evaluation. To measure whether a targeted source sentence is excluded from the generated summary, we compute Sentence-BERT (Reimers and Gurevych, 2019) cosine similarity between the targeted sentence and each generated summary sentence, using the maximum similarity as the alignment score. A sentence

is considered included if this score exceeds a fixed threshold and excluded otherwise. We report lead sentence inclusion as the percentage of cases where perturbations fail to exclude lead content; the same threshold is used across all models and datasets, and moderate threshold changes do not alter the relative trends. We use a threshold of 0.1 across all models and datasets and verify that moderate threshold changes (0.05–0.30) do not alter the relative trends. **Summarization Quality.** We evaluate summarization quality before and after BA-TVN using ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) for lexical overlap, BERTScore (Zhang et al.) for semantic similarity, BARTScore (Yuan et al., 2021) for generation quality, and SummaC (Laban et al., 2022) and FactCC (Kryściński et al., 2020) for factual consistency. These metrics verify that reducing positional reliance does not introduce semantic drift or factual errors that ROUGE alone may miss. We assess before-and-after quality changes using Wilcoxon signed-rank tests.

Defense-aware adaptive adversaries In addition to standard perturbations targeting lead positions, we evaluate robustness under adaptive defense-aware perturbations applied to non-lead positions (sentences 4-10). This setting captures scenarios where perturbations are shifted away from lead content after observing that lead-focused attacks are less effective, applying the same character-, word-, sentence-, and document-level modifications outside the lead region.

6 Results and Discussion

6.1 Evaluation of Lead Bias Mitigation

We evaluate whether BA-TVN reduces excessive lead-position reliance across eleven models and six datasets using Lead Bias Ratio (LBR), which measures the proportion of summary content derived from the first three sentences relative to the full document. Table 1 shows that BA-TVN consistently lowers LBR, achieving a stable average reduction of 31% across model–dataset conditions, with a bootstrap 95% confidence interval of [29.4, 32.7]%. Reductions are strongest on news datasets, while scientific datasets retain moderately higher post-mitigation LBR because introductory content often carries genuine informational value. Percentage-based lead windows covering the first 10–20% of each document yield similar reductions of 32–35%, confirming that the trend is not tied to the fixed three-sentence window. Overall, TVN reduces excessive positional reliance without eliminating useful discourse structure.

6.2 Robustness Against Sentence Exclusion

We evaluate whether BA-TVN improves robustness against sentence exclusion attacks that exploit lead-bias vulnerabilities. Following (Thota and Nilizadeh, 2024), we apply character-, word-, sentence-, and document-level perturbations to lead sentences and measure *lead sentence inclusion*, the percentage of summaries that still retain content from the targeted lead sentence. Lower inclusion indicates more successful sentence exclusion. Table 2 reports inclusion before attack, after attack on undefended models, and after attack on BA-TVN-defended models. Undefended models show a large drop in inclusion under attack, confirming that the perturbations cause meaningful content omission. After applying BA-TVN, models retain substantially higher inclusion under the same attacks, recovering roughly 25–28 percentage points across datasets and model families, with an average gain of 27.3 points and a bootstrap 95% confidence interval of [25.8, 28.6]. Full per-perturbation results are provided in Appendix 10.3.1.

6.3 Summarization Quality Preservation

A critical concern with reducing positional reliance is whether the edit degrades summarization quality. To evaluate this, we measure lexical overlap, semantic similarity, generation quality, and factual consistency before and after applying TVN with $\lambda = 0.3$. Table 3 reports results on Multi-News as a representative dataset. Across model families, ROUGE-1 changes remain small, typically within 0.003, and BERTScore, BARTScore, SummaC, and FactCC show similarly minor variation. Wilcoxon signed-rank tests show that the before–after differences are not statistically significant ($p > 0.05$), indicating that BA-TVN does not introduce measurable degradation in summarization quality. The full per-dataset results in Appendix 10.3.2 show the same pattern with only minor metric changes across all evaluated models.

6.4 Ablation Studies

We conduct ablation studies to assess the sensitivity of TVN to key design choices and dataset characteristics. Overall, these analyses indicate that TVN is robust to moderate variations in hyperparameters and bias construction strategies. ***Scaling coefficient.*** We vary the task-vector scaling coefficient λ

Table 1: LBR before and after BA-TVN across models and datasets. Lower LBR indicates reduced positional shortcut reliance. BA-TVN achieves an average LBR reduction of 31% with 95% CI [29.4, 32.7]%.

Model	Datasets											
	MultiNews		CNN/DM		WikiSum		ArXiv		PubMed		Multi-XScience	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
BART-Large	0.87	0.58	0.84	0.57	0.86	0.60	0.85	0.59	0.83	0.56	0.89	0.67
T5-Small	0.82	0.61	0.80	0.59	0.83	0.62	0.81	0.60	0.79	0.58	0.91	0.71
Pegasus-Large	0.82	0.56	0.81	0.55	0.84	0.58	0.82	0.57	0.80	0.54	0.86	0.64
Flan-T5-Large	0.81	0.44	0.79	0.42	0.82	0.45	0.80	0.43	0.78	0.40	0.85	0.52
Qwen2.5B	0.79	0.46	0.77	0.44	0.80	0.47	0.78	0.45	0.76	0.42	0.83	0.55
Qwen-3	0.78	0.44	0.76	0.42	0.79	0.45	0.77	0.43	0.75	0.40	0.82	0.53
Mistral-7B	0.80	0.47	0.78	0.45	0.81	0.48	0.79	0.46	0.77	0.43	0.84	0.56
Vicuna-7B	0.81	0.48	0.79	0.46	0.82	0.49	0.80	0.47	0.78	0.44	0.85	0.57
LLaMA-3-8B	0.79	0.45	0.77	0.43	0.80	0.46	0.78	0.44	0.76	0.41	0.83	0.55
Gemma-3-it	0.80	0.43	0.78	0.41	0.81	0.44	0.79	0.42	0.77	0.39	0.84	0.54
GPT-oss-20b	0.81	0.42	0.79	0.40	0.82	0.43	0.80	0.41	0.78	0.38	0.85	0.53

Table 2: Lead sentence inclusion under adversarial perturbations. BA-TVN Gain denotes recovery over undefended models; higher values indicate better retention. Mean gain: 27.3 points, 95% CI [25.8, 28.6].

Model	MultiNews				Multi-XScience			
	Clean	Undef. Attack	BA-TVN Attack	BA-TVN Gain	Clean	Undef. Attack	BA-TVN Attack	BA-TVN Gain
BART-Large	87.4	42.8	**69.4**	26.6	73.2	38.9	**64.8**	25.9
T5-Small	82.6	39.5	**66.9**	27.4	70.0	36.1	**62.3**	26.2
Pegasus-Large	82.7	40.2	**67.8**	27.6	71.5	36.8	**63.1**	26.3
Flan-T5-Large	72.8	37.6	**65.2**	27.6	69.0	34.9	**60.6**	25.7
LLaMA-3-8B	83.4	47.5	**75.8**	28.3	78.9	43.7	**70.8**	27.1
Qwen2.5B	81.6	45.8	**74.0**	28.2	77.6	42.1	**69.6**	27.5
Qwen-3	82.1	46.3	**74.7**	28.4	78.1	42.7	**70.1**	27.4
Mistral-7B	80.7	44.9	**73.3**	28.4	76.8	41.4	**68.8**	27.4
Vicuna-7B	79.9	44.1	**72.5**	28.4	76.1	40.8	**68.1**	27.3
Gemma-3-it	82.8	47.1	**75.2**	28.1	79.4	43.2	**70.5**	27.3
GPT-oss-20b	83.5	48.0	**76.0**	28.0	80.1	44.0	**71.2**	27.2

Table 3: Summarization quality before and after BA-TVN on MultiNews. Higher is better for ROUGE-1, BERTScore, SummaC, and FactCC. For BARTScore, less negative values indicate better generation quality.

Model	R-1		BERT		BART		SummaC		FactCC	
	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.
BART-Large	0.325	0.327	0.879	0.881	-2.430	-2.428	0.718	0.719	0.884	0.883
T5-Small	0.410	0.412	0.885	0.887	-2.405	-2.402	0.724	0.725	0.889	0.888
Pegasus-Large	0.370	0.369	0.882	0.883	-2.418	-2.417	0.721	0.722	0.887	0.886
Flan-T5-Large	0.418	0.420	0.889	0.891	-2.392	-2.390	0.729	0.730	0.894	0.893
Qwen2.5B	0.402	0.404	0.886	0.888	-2.401	-2.399	0.726	0.727	0.891	0.890
Qwen-3	0.408	0.410	0.887	0.889	-2.397	-2.395	0.727	0.728	0.892	0.891
Mistral-7B	0.395	0.397	0.884	0.886	-2.410	-2.408	0.724	0.725	0.889	0.888
Vicuna-7B	0.392	0.394	0.883	0.885	-2.414	-2.411	0.723	0.724	0.888	0.887
LLaMA-3-8B	0.415	0.417	0.890	0.892	-2.386	-2.384	0.731	0.732	0.895	0.894
Gemma-3-it	0.421	0.422	0.891	0.893	-2.382	-2.380	0.732	0.733	0.896	0.895
GPT-oss-20b	0.424	0.425	0.893	0.894	-2.376	-2.374	0.734	0.735	0.897	0.896

to study trade-off between LBR reduction and summarization quality. Smaller values ($\lambda = 0.1, 0.2$) yield limited LBR reductions (8–15%), while larger values ($\lambda = 0.5, 0.7$) achieve stronger reductions (35–40%) but cause ROUGE degradation (5–8%). Across models and datasets, $\lambda = 0.3$ provides the best balance between positional-bias reduction and quality preservation. ***Bias amplification strength.*** We also vary the fraction of bias-amplified samples from 25% to 100% of the fine-tuning data. At 25%, the positional signal is often too weak for reliable task-vector extraction, leading to inconsistent debiasing. At 50% and above, the signal stabilizes and LBR reductions remain consistent, suggesting that TVN is not highly sensitive to the exact amplification fraction once sufficient bias concentration is reached. ***Bias amplification strategies.*** Finally, we compare document-based amplification, which moves summary-relevant content toward lead positions, with summary-based amplification, which modifies references to emphasize lead content. Both strategies achieve 90–95% bias concentration and produce similar debiasing trends. Document restructuring yields slightly stronger signals for heterogeneous documents and is used in our primary experiments, while summary modification is more computationally efficient.

6.5 Comparison with Bias Mitigation

We compare BA-TVN against established lead-bias mitigation baselines, including multi-stage training, AdvSumm, and NeuS, which target positional or framing bias through training- or generation-time interventions. Table 4 reports results on MultiNews and Multi-XScience, with full results for all eleven models and six datasets in Appendix 10.3.3. Across both datasets, BA-TVN achieves substantially lower LBR than prior baselines, reducing LBR to approximately 0.40–0.55 on most models, while competing methods often remain above 0.70, especially on long-form scientific data. These reductions align with the robustness gains in Section 6.2, suggesting that reducing positional shortcut reliance weakens perturbation-based sentence exclusion.

6.6 Comparison with Unlearning Strategies

We compare BA-TVN against generic machine unlearning baselines, including approximate and exact unlearning. Unlike BA-TVN, these methods are not designed for positional bias mitigation; they use the bias-amplified data only as unlearning targets and then apply their own optimization procedures.

Table 4: Comparison with existing lead-bias mitigation baselines using LBR. Lower values indicate reduced positional shortcut reliance.

Model	MultiNews					Multi-XScience				
	Orig	MS	Adv	NeuS	BA-TVN	Orig	MS	Adv	NeuS	BA-TVN
BART-Large	0.87	0.79	0.77	0.74	**0.58**	0.89	0.84	0.82	0.80	**0.67**
T5-Small	0.82	0.76	0.74	0.72	**0.61**	0.91	0.86	0.84	0.82	**0.71**
Pegasus-Large	0.82	0.77	0.75	0.73	**0.56**	0.86	0.82	0.80	0.78	**0.64**
Flan-T5-Large	0.81	0.78	0.76	0.74	**0.44**	0.85	0.81	0.79	0.77	**0.52**
Qwen2.5B	0.79	0.77	0.75	0.74	**0.46**	0.83	0.80	0.79	0.77	**0.55**
Qwen-3	0.78	0.76	0.74	0.73	**0.44**	0.82	0.79	0.78	0.76	**0.53**
Mistral-7B	0.80	0.78	0.76	0.74	**0.47**	0.84	0.81	0.80	0.78	**0.56**
Vicuna-7B	0.81	0.79	0.77	0.75	**0.48**	0.85	0.82	0.81	0.79	**0.57**
LLaMA-3-8B	0.79	0.77	0.75	0.73	**0.45**	0.83	0.80	0.79	0.77	**0.55**
gemma-3-it	0.80	0.78	0.76	0.74	**0.43**	0.84	0.81	0.80	0.78	**0.54**
gpt-oss-20b	0.81	0.79	0.77	0.75	**0.42**	0.85	0.82	0.81	0.79	**0.53**

Table 5: Comparison with generic unlearning strategies using LBR. Lower values indicate reduced positional shortcut reliance.

Model	MultiNews				Multi-XScience			
	Orig	Approx	Exact	BA-TVN	Orig	Approx	Exact	BA-TVN
BART-Large	0.87	0.76	0.73	**0.58**	0.89	0.81	0.78	**0.67**
T5-Small	0.82	0.74	0.71	**0.61**	0.91	0.84	0.81	**0.71**
Pegasus-Large	0.82	0.75	0.72	**0.56**	0.86	0.80	0.77	**0.64**
Flan-T5-Large	0.81	0.73	0.70	**0.44**	0.85	0.79	0.76	**0.52**
Qwen2.5B	0.79	0.72	0.69	**0.46**	0.83	0.77	0.74	**0.55**
Qwen-3	0.78	0.71	0.68	**0.44**	0.82	0.76	0.73	**0.53**
Mistral-7B	0.80	0.73	0.70	**0.47**	0.84	0.78	0.75	**0.56**
Vicuna-7B	0.81	0.74	0.71	**0.48**	0.85	0.79	0.76	**0.57**
LLaMA-3-8B	0.79	0.72	0.69	**0.45**	0.83	0.77	0.74	**0.55**
gemma-3-it	0.80	0.72	0.68	**0.43**	0.84	0.77	0.73	**0.54**
gpt-oss-20b	0.81	0.73	0.69	**0.42**	0.85	0.78	0.74	**0.53**

Table 5 reports results on MultiNews and Multi-XScience, with full results across all six datasets in Appendix 10.3.4. Both unlearning baselines reduce LBR, with exact unlearning generally outperforming approximate unlearning. However, their reductions remain limited, especially on Multi-XScience, where LBR often stays above 0.75. In contrast, BA-TVN achieves consistently lower LBR across model families. Appendix results show the same trend across the remaining datasets, suggesting that directly suppressing the positional shortcut direction is more effective than generic unlearning.

6.7 Comparison with Robustness Defenses

Table 6 reports LBR on MultiNews and Multi-XScience, with full results across all datasets in Appendix 10.3.5. Input-level defenses provide limited LBR reductions because they primarily normalize or reconstruct perturbed tokens without changing the model's reliance on lead-position information. Adversarial data augmentation performs better than purification-based defenses, but LBR remains relatively high. In contrast, BA-TVN consistently achieves the lowest LBR across different model families. These results suggest that directly editing positional shortcut behavior in parameter space is more effective than treating perturbations only as input-level artifacts.

Table 6: Comparison with robustness-oriented defenses using LBR. Lower values indicate reduced positional shortcut reliance.

Model	MultiNews					Multi-XScience				
	Purif.	scRNN	Lang.	AdvAug	BA-TVN	Purif.	scRNN	Lang.	AdvAug	BA-TVN
BART-Large	0.82	0.84	0.83	0.71	**0.58**	0.85	0.86	0.85	0.76	**0.67**
T5-Small	0.78	0.79	0.78	0.69	**0.61**	0.87	0.88	0.87	0.78	**0.71**
Pegasus-Large	0.77	0.78	0.78	0.68	**0.56**	0.82	0.83	0.82	0.74	**0.64**
Flan-T5-Large	0.76	0.77	0.76	0.66	**0.44**	0.81	0.82	0.81	0.72	**0.52**
Qwen2.5B	0.74	0.75	0.74	0.64	**0.46**	0.79	0.80	0.79	0.70	**0.55**
Qwen-3	0.73	0.74	0.73	0.63	**0.44**	0.78	0.79	0.78	0.69	**0.53**
Mistral-7B	0.75	0.76	0.75	0.65	**0.47**	0.80	0.81	0.80	0.71	**0.56**
Vicuna-7B	0.76	0.77	0.76	0.66	**0.48**	0.81	0.82	0.81	0.72	**0.57**
LLaMA-3-8B	0.74	0.75	0.74	0.64	**0.45**	0.79	0.80	0.79	0.70	**0.55**
gemma-3-it	0.74	0.75	0.74	0.63	**0.43**	0.80	0.81	0.80	0.70	**0.54**
gpt-oss-20b	0.75	0.76	0.75	0.63	**0.42**	0.81	0.82	0.81	0.71	**0.53**

6.8 Robustness to Adaptive Attack Strategies

We evaluate BA-TVN under adaptive attacks, where perturbations target non-lead sentences (4–10) instead of lead sentences. On undefended models, these perturbations cause modest ROUGE-1 drops of 3–7 points, indicating that positional shortcut reliance amplifies sensitivity beyond the lead region. After applying TVN, ROUGE-1 recovers to within 1 point of the clean baseline across all models and datasets, and generated summaries remain semantically coherent. These results show that reducing reliance on positional shortcuts generalizes beyond lead content, with minimal impact on summary generation under non-lead perturbations.

7 Conclusion

In this work, we investigated whether parameter-space editing can mitigate positional shortcut reliance in text summarization models and improve robustness to perturbation-based sentence exclusion attacks. We introduced Bias-Amplified Task Vector Negation (BA-TVN), which isolates excessive lead-position reliance through bias-amplified summarization signals and suppresses it via scaled vector negation. Across six datasets and eleven model architectures, BA-TVN consistently reduces Lead Bias Ratio (LBR), improves robustness against sentence exclusion attacks, and preserves summarization quality across lexical, semantic, generation-quality, and factual consistency metrics. Compared to prior bias-mitigation, generic unlearning, and input-level robustness baselines, BA-TVN achieves stronger and more stable reductions in positional reliance. These findings suggest that bias-amplified parameter-space editing is a practical and architecture-agnostic approach for mitigating positional shortcut reliance in summarization models.

8 Limitations

Our approach assumes access to model parameters during development in order to construct bias-amplified training data, perform fine-tuning, and apply task-vector editing. Accordingly, BA-TVN is primarily applicable to open-weight summarization models and enterprise deployment settings where models are fine-tuned or adapted internally. Deployments relying exclusively on closed-source API-based systems fall outside the scope of this work, since parameter-space editing is not directly feasible without access to model weights or adaptation layers. In addition, our evaluation focuses specifically on perturbation-based sentence exclusion behavior associated with positional shortcut reliance in abstractive summarization models. We do not claim robustness against all forms of adversarial manipulation or broader summarization failure modes such as hallucination or reasoning errors unrelated to positional bias. Future work can extend BA-TVN to conversational summarization, broader structural shortcuts, and lightweight adaptation for closed-source systems, while complementing LBR with human coverage judgments.

9 Ethical Considerations

This work studies adversarial vulnerabilities and defense mechanisms for text summarization models. A potential risk of this work is that the perturbation strategies analyzed in our evaluation could be misused to manipulate summarization outputs or suppress important information. To mitigate this risk, we present these perturbations only as controlled evaluation tools for studying robustness, and we focus on developing a defense that reduces model sensitivity to such manipulations. All adversarial perturbations evaluated in this paper are applied to publicly available benchmark datasets and are intended solely to assess model robustness under controlled experimental settings. We do not deploy, encourage, or facilitate malicious misuse of summarization systems. Our proposed defense operates during model development and does not involve monitoring user behavior, collecting personal data, or modifying user inputs at inference time. The evaluation datasets used in this study are standard benchmarks that do not contain personally identifiable information. By improving robustness to adversarial perturbations, our work aims to enhance the reliability and trustworthiness of summarization systems deployed in high-stakes domains such as policy analysis, scientific communication, and information aggregation. We believe that proactively addressing structural vulnerabilities contributes positively to the safe and responsible deployment of generative language models.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthropic. 2025. Introducing claude 4.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-jun Li, and Yaxing Yao. 2025. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 277–297.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621.

Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. Wikisum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 212–219.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Omkar Dige, Diljot Arneja, Tsz Fung Yau, Qixuan Zhang, Mohammad Bolandraftar, Xiaodan Zhu, and Faiza Khattak. 2024. Can machine unlearning reduce social bias in language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 954–969.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Alexander Richard 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.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. *arXiv preprint arXiv:1909.04028*.

Mukur Gupta, Nikhil Reddy Varimalla, Nicholas Deas, Melanie Subbiah, and Kathleen McKeown. 2025. Advsumm: Adversarial training for bias mitigation in text summarization. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 172–182.

Kung-Hsiang Huang, Philippe Laban, Alexander Richard Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. 2021. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pages 2008–2016. PMLR.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de Las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, and 1 others. 2023. Mistral 7b. corr, abs/2310.06825, 2023. doi: 10.48550. *arXiv preprint ARXIV.2310.06825*, 10.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4.

Chris Kedzie, Kathleen Mckeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

Wojciech Kryściński, 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.

Philippe Laban, Tobias Schnabel, Paul 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.

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.

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. *arXiv preprint arXiv:1910.13461*.

Linyang Li, Demin Song, and Xipeng Qiu. 2023. Text adversarial purification as defense against adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 338–350.

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.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Adam Makepeace and contributors. 2021. language-tool-python: A wrapper for languagetool. https://github.com/languagetool-org/language-tool-python.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.

Raha Moraffah, Shubh Khandelwal, Amrita Bhattacharjee, and Huan Liu. 2024. Adversarial text purification: A large language model approach for defense. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 65–77.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

OpenAI. 2025. Introducing gpt-5.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591.

A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint*.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Ronald Rosenfeld and 1 others. 1996. A maximum entropy approach to adaptive statistical language modelling. *Computer speech and language*, 10(3):187.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Lingfeng Shen, Ze Zhang, Haiyun Jiang, and Ying Chen. Textshield: Beyond successfully detecting adversarial sentences in text classification. In *The Eleventh International Conference on Learning Representations*.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1569–1576.

Julius Steen and Katja Markert. 2024. Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5962–5983.

Poojitha Thota and Shirin Nilizadeh. 2024. Attacks against abstractive text summarization models through lead bias and influence functions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13727–13741.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In

2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), pages 303–319. IEEE.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

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, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1462–1471.

10 Appendix

10.1 Example showing the impact of adversarial perturbations

Following prior work (Thota and Nilizadeh, 2024), we generate adversarial samples using established perturbation techniques to demonstrate their impact on text summarization models. Table 7 presents an example of a word-level homoglyph attack, showing the dramatic effect of subtle character modifications on summary generation. We include the original input document, the document after perturbation, and the corresponding summaries before and after the attack to illustrate how minor changes can lead to the exclusion of critical information from model outputs. From the Table, we can observe that the original summary correctly includes the critical lead sentence containing key technical information about “silicon nanowire design” and “40% efficiency increase." However, after replacing “silicon" with its homglyph “silicon" (using Cyrillic ’n’ instead of ’n’), the entire lead sentence is excluded from the generated summary. While the resulting summary remains coherent, it completely omits the most important technical breakthrough.

10.2 Scaling coefficient and quality trade-off.

We also vary the task-vector scaling coefficient λ to analyze the trade-off between positional shortcut reduction and summarization quality. Smaller values ($\lambda = 0.1, 0.2$) preserve quality but yield limited LBR reductions (8–15%). Larger values ($\lambda = 0.5, 0.7$) reduce LBR more strongly (35–40%) but introduce noticeable quality degradation, including ROUGE drops of 5–8% and small declines in factuality metrics. Across models and datasets, $\lambda = 0.3$ lies near the best trade-off region: it achieves substantial LBR reduction while keeping ROUGE, BERTScore, SummaC, and FactCC stable. This suggests that TVN can be tuned along a bias–quality trade-off curve, with $\lambda = 0.3$ providing the most favorable balance in our experiments.

10.3 Additional Results

10.3.1 Per-Perturbation Robustness Results

In the main paper, we report lead sentence inclusion before attack, after attack on undefended models, and after attack on TVN defended models. Here we present per-perturbation results in detail for all six datasets using eleven summarization models. We follow the same measurement protocol as (Thota and Nilizadeh, 2024). Specifically, we apply character-level, word-level, sentence-level, and document-level perturbations to the lead sentence and report *lead sentence inclusion*, defined as the percentage of summaries that retain semantic content from the original lead sentence after perturbation. Higher values indicate that perturbations fail to exclude the lead sentence from the generated summary. As shown in Table 8, TVN consistently improves robustness across all perturbation types and datasets. While sentence- and document-level attacks (e.g., SRH, SRP, DR) induce moderate degradation, the severe exclusion effects observed in prior attack-only settings are no longer present. These trends hold across both news-style datasets and long-form scientific datasets, indicating that reducing positional dependence at the parameter level weakens the core mechanism exploited by sentence exclusion attacks.

10.3.2 Detailed Summarization Quality Results

We report summarization quality on MultiNews as a representative dataset in Section 6.3. Here, we provide the corresponding quality results for the remaining five datasets: CNN/DailyMail, Wik-

Table 7: Summary before and after Character Replacement with Homoglyph

Element	Description
Input Document	Scientists at Stanford University announced a breakthrough in renewable energy technology that could revolutionize solar power efficiency. The new **silicon** nanowire design increases energy conversion rates by 40% compared to traditional panels. Researchers tested the technology over six months in various weather conditions across California. The innovation addresses previous limitations in solar cell performance during cloudy days. "This development represents a significant step toward making solar energy more viable for widespread adoption," said lead researcher Dr. Sarah Chen. The team plans to begin commercial trials next year.
Original Summary	Stanford scientists announced breakthrough in renewable energy with new silicon nanowire design that increases solar power efficiency by 40%. Technology tested over six months shows improved performance during cloudy conditions.
Input Document after Perturbation	Scientists at Stanford University announced a breakthrough in renewable energy technology that could revolutionize solar power efficiency. The new **silicon** nanowire design increases energy conversion rates by 40% compared to traditional panels. Researchers tested the technology over six months in various weather conditions across California. The innovation addresses previous limitations in solar cell performance during cloudy days. "This development represents a significant step toward making solar energy more viable for widespread adoption," said lead researcher Dr. Sarah Chen. The team plans to begin commercial trials next year.
Summary after Perturbation	Technology tested over six months shows improved performance during cloudy conditions. Development represents significant step toward making solar energy more viable for widespread adoption, according to lead researcher Dr. Sarah Chen.

iSum, ArXiv, PubMed, and Multi-XScience. Table 9 reports ROUGE-1, BERTScore, BARTScore, SummaC, and FactCC before and after applying TVN across all eleven summarization models. Across datasets and model families, the before-and-after changes remain small. ROUGE-1 varies only marginally after TVN, while BERTScore, BARTScore, SummaC, and FactCC remain stable across encoder-decoder models, decoder-only models, and recent open-weight models such as Gemma-3-it and GPT-oss-20b. These results indicate that reducing positional shortcut reliance does not introduce measurable degradation in lexical overlap, semantic coherence, generation quality, or factual consistency.

10.3.3 Detailed Bias-Mitigation Baseline Results

We show a comparison of TVN with lead-bias mitigation baselines on MultiNews and Multi-XScience in Section 6.7. Here, we provide the corresponding results for the remaining four datasets: CNN/DailyMail, WikiSum, ArXiv, and PubMed. Table 10 reports LBR for the original model, multi-stage training (MS), AdvSumm (Adv), NeuS, and TVN across all eleven summarization models. Lower LBR indicates reduced positional shortcut reliance. Across all remaining datasets, the trends are consistent with the results obtained with the MultiNews and Multi-Xscience datasets. Existing bias-mitigation baselines reduce LBR moderately compared to the original models, but they often retain substantial positional reliance. In contrast, TVN achieves the lowest LBR across datasets and model families, including recent open-weight models such as Gemma-3-it and GPT-oss-20b. These results indicate that the stronger reductions observed are not specific to MultiNews or Multi-XScience, but generalize across news, encyclopedic, and scientific summarization settings.

10.3.4 Detailed Unlearning Baseline Results

We compare TVN with generic unlearning baselines on MultiNews and Multi-XScience, shown in Section 6.6. Here, we provide the corresponding results for the remaining four datasets: CNN/DailyMail, WikiSum, ArXiv, and PubMed. Table 11 reports LBR for the original model, approximate unlearning, exact unlearning, and TVN across all eleven summarization models. Across all remaining datasets, both unlearning baselines reduce LBR compared to the original models, with exact unlearning generally performing better than approximate unlearning. However, the reductions remain smaller than those achieved by TVN. This trend is consistent across encoder-decoder models, decoder-only models, and recent open-weight models such as Gemma-3-it and GPT-oss-20b, indicating that generic unlearning strategies do not suppress positional shortcut reliance as effectively as BA-TVN.

10.3.5 Detailed Robustness-Defense Baseline Results

In the main paper, we compare TVN with robustness-oriented defenses on MultiNews and Multi-XScience in Section 6.7. Here, we provide the corresponding results for the remaining four datasets: CNN/DailyMail, WikiSum, ArXiv, and PubMed. Table 12 reports LBR for text adversarial purification, scRNN, LanguageTool normalization, adversarial data augmentation, and TVN across all eleven summarization models. Across all remaining datasets, input-level defenses reduce LBR only modestly because they primarily normalize perturbations rather than changing the model's positional reliance. Adversarial data augmentation provides stronger reductions than purification-based defenses, but it still leaves higher LBR values than TVN. These trends are consistent across encoder-decoder models, decoder-only models, and recent open-weight models such as Gemma-3-it and GPT-oss-20b, indicating that directly editing the positional shortcut in parameter space is more effective than treating perturbations only at the input or training-data level.

Table 8: Percentage of lead sentence inclusion before and after adversarial perturbations *after applying TVN* on additional datasets. Higher values indicate that perturbations fail to exclude the lead sentence from the summary.

Dataset	**Model**	**Before**	**After Perturbation (post-TVN)**										
			CI	CD	CR	CS	WD	WRS	WRH	SR	SRH	SRP	DR
MultiNews	BART-Large	87.4	75.6	76.1	72.4	74.8	71.2	73.9	70.6	66.8	60.7	62.3	58.9
	T5-Small	82.6	72.9	73.4	69.7	72.1	68.6	71.2	67.9	64.2	58.5	60.1	56.8
	Pegasus-Large	82.7	73.8	74.1	70.5	73.0	69.4	72.6	68.8	65.5	59.8	61.0	57.6
	Flan-T5-Large	72.8	70.9	71.3	68.0	70.6	66.9	69.7	66.1	62.7	57.2	58.4	55.0
	LLaMA-3-8B	83.4	81.0	81.6	78.8	80.9	77.4	79.6	76.8	73.1	68.7	69.9	66.4
	Qwen2.5B	81.6	79.2	79.8	76.9	79.0	75.6	77.9	75.1	71.4	66.8	68.0	64.6
	Qwen-3	82.1	79.8	80.4	77.6	79.7	76.3	78.6	75.8	72.2	67.6	68.8	65.4
	Mistral-7B	80.7	78.4	79.0	76.1	78.2	74.8	77.2	74.4	70.9	66.1	67.4	64.0
	Vicuna-7B	79.9	77.6	78.1	75.3	77.4	74.0	76.3	73.6	70.1	65.4	66.6	63.2
	Gemma-3-it	82.8	80.6	81.2	78.3	80.5	77.0	79.2	76.4	72.8	68.2	69.5	66.0
	GPT-oss-20b	83.5	81.4	82.0	79.1	81.3	77.8	80.0	77.2	73.6	69.0	70.3	66.8
Multi-XScience	BART-Large	73.2	70.8	71.4	67.9	70.6	66.5	69.7	65.9	62.1	56.4	57.9	53.8
	T5-Small	70	68.4	68.9	65.3	68.1	64.0	67.1	63.2	59.5	54.1	55.4	51.6
	Pegasus-Large	71.5	69.1	69.6	66.1	68.9	64.8	67.9	64.0	60.3	55.0	56.2	52.4
	Flan-T5-Large	69.0	66.2	66.8	63.4	66.0	62.1	65.1	61.6	58.0	53.0	54.2	50.6
	LLaMA-3-8B	78.9	76.1	76.8	73.9	75.9	72.3	74.6	71.8	68.4	63.5	64.7	60.9
	Qwen2.5B	77.6	74.9	75.5	72.7	74.6	71.2	73.4	70.6	67.2	62.3	63.6	59.8
	Qwen-3	78.1	75.3	76.0	73.1	75.0	71.6	73.9	71.1	67.8	62.9	64.1	60.3
	Mistral-7B	76.8	74.1	74.7	71.8	73.7	70.3	72.6	69.9	66.5	61.6	62.8	59.2
	Vicuna-7B	76.1	73.4	74.0	71.1	73.0	69.7	71.9	69.2	65.8	60.9	62.1	58.4
	Gemma-3-it	79.4	76.5	77.2	74.3	76.3	72.8	75.1	72.3	68.9	64.0	65.2	61.4
	GPT-oss-20b	80.1	77.2	77.9	75.0	77.0	73.5	75.8	73.0	69.6	64.7	65.9	62.1
CNN/DailyMail	BART-Large	84.1	73.2	73.8	70.4	72.6	69.1	71.9	68.4	64.7	59.2	60.8	56.5
	T5-Small	81.6	71.4	72.0	68.5	70.7	67.2	69.8	66.4	62.8	57.4	58.9	54.6
	Pegasus-Large	83.5	74.0	74.6	71.1	73.3	69.8	72.7	69.1	65.4	60.1	61.6	57.2
	Flan-T5-Large	79.8	71.9	72.5	69.0	71.3	67.8	70.4	67.1	63.5	58.2	59.7	55.3
	LLaMA-3-8B	82.7	80.4	81.0	78.1	80.2	76.8	79.0	76.1	72.5	68.0	69.3	65.8
	Qwen-2.5B	81.9	79.3	79.9	77.0	79.1	75.7	77.9	75.1	71.4	66.8	68.0	64.6
	Qwen-3	82.3	79.8	80.4	77.6	79.7	76.3	78.6	75.8	72.2	67.6	68.8	65.4
	Mistral-7B	80.7	78.4	79.0	76.1	78.2	74.8	77.2	74.4	70.9	66.1	67.4	64.0
	Vicuna-7B	79.9	77.6	78.1	75.3	77.4	74.0	76.3	73.6	70.1	65.4	66.6	63.2
	Gemma-3-it	82.9	80.7	81.3	78.4	80.5	77.1	79.3	76.5	73.0	68.4	69.7	66.2
	GPT-oss-20b	83.6	81.5	82.1	79.2	81.3	77.9	80.1	77.3	73.8	69.2	70.5	67.0
WikiSum	BART-Large	86.2	75.9	76.4	72.8	75.1	71.6	74.2	70.8	67.2	61.9	63.4	59.1
	T5-Small	83.7	74.2	74.8	71.2	73.5	70.0	72.6	69.2	65.6	60.3	61.8	57.6
	Pegasus-Large	85.7	76.6	77.1	73.6	75.9	72.3	75.0	71.6	67.9	62.6	64.1	59.8
	Flan-T5-Large	82.9	74.8	75.3	71.8	74.1	70.6	73.1	69.8	66.1	60.8	62.3	58.1
	LLaMA-3-8B	84.1	81.6	82.2	79.3	81.4	78.0	80.2	77.4	73.9	69.3	70.7	67.1
	Qwen-2.5B	83.5	80.7	81.3	78.5	80.6	77.2	79.4	76.6	73.0	68.5	69.9	66.3
	Qwen-3	83.9	81.1	81.7	78.9	81.0	77.6	79.8	77.0	73.5	68.9	70.3	66.8
	Mistral-7B	83.4	80.8	81.3	78.5	80.6	77.2	79.4	76.6	73.0	68.5	69.9	66.3
	Vicuna-7B	82.6	79.9	80.5	77.7	79.8	76.4	78.6	75.9	72.3	67.8	69.1	65.6
	Gemma-3-it	84.5	82.1	82.7	79.8	81.9	78.5	80.7	77.9	74.4	69.8	71.2	67.6
	GPT-oss-20b	85.1	82.8	83.4	80.5	82.6	79.2	81.4	78.6	75.1	70.5	71.9	68.3
ArXiv	BART-Large	82.8	74.5	75.1	71.6	73.9	70.4	72.8	69.3	65.7	60.9	62.1	58.4
	T5-Small	80.4	72.7	73.3	69.8	72.1	68.6	71.0	67.5	63.9	59.1	60.3	56.6
	Pegasus-Large	83.1	75.2	75.8	72.3	74.6	71.1	73.6	70.1	66.4	61.7	62.9	59.2
	Flan-T5-Large	80.9	73.3	73.9	70.4	72.7	69.2	71.6	68.1	64.5	59.7	60.9	57.2
	LLaMA-3-8B	82.3	80.1	80.7	77.8	79.9	76.6	78.8	76.0	72.4	67.9	69.2	65.6
	Qwen-2.5B	81.7	79.3	79.9	77.0	79.1	75.7	77.9	75.1	71.4	66.8	68.1	64.7
	Qwen-3	82.0	79.7	80.3	77.4	79.5	76.1	78.3	75.5	71.9	67.3	68.6	65.1
	Mistral-7B	81.5	79.0	79.6	76.7	78.8	75.4	77.6	74.9	71.3	66.8	68.1	64.7
	Vicuna-7B	80.8	78.2	78.8	75.9	78.0	74.6	76.8	74.0	70.4	65.9	67.2	63.7
	Gemma-3-it	82.7	80.4	81.0	78.1	80.2	76.9	79.1	76.3	72.8	68.3	69.6	66.0
	GPT-oss-20b	83.4	81.1	81.7	78.8	80.9	77.6	79.8	77.0	73.5	69.0	70.3	66.7
PubMed	BART-Large	81.9	74.8	75.4	71.9	74.2	70.7	73.1	69.6	66.1	61.4	62.7	58.9
	T5-Small	79.6	73.1	73.7	70.2	72.5	69.0	71.4	67.9	64.3	59.6	60.8	57.0
	Pegasus-Large	82.4	75.5	76.1	72.6	74.9	71.4	73.9	70.3	66.8	62.1	63.4	59.6
	Flan-T5-Large	80.1	73.7	74.3	70.8	73.1	69.6	72.0	68.6	65.0	60.3	61.6	57.8
	LLaMA-3-8B	81.6	79.7	80.3	77.4	79.5	76.2	78.4	75.6	72.0	67.5	68.8	65.2
	Qwen-2.5B	81.0	78.9	79.5	76.6	78.7	75.3	77.5	74.8	71.1	66.6	67.9	64.3
	Qwen-3	81.3	79.3	79.9	77.0	79.1	75.7	77.9	75.1	71.4	66.9	68.2	64.6
	Mistral-7B	80.8	78.6	79.2	76.3	78.4	75.0	77.2	74.4	70.8	66.3	67.6	64.1
	Vicuna-7B	80.0	77.8	78.4	75.5	77.6	74.2	76.4	73.7	70.0	65.5	66.8	63.3
	Gemma-3-it	82.0	80.0	80.6	77.7	79.8	76.5	78.7	75.9	72.3	67.8	69.1	65.5
	GPT-oss-20b	82.7	80.7	81.3	78.4	80.5	77.2	79.4	76.6	73.0	68.5	69.8	66.2

Table 9: Summarization quality before and after TVN on the remaining datasets. Higher is better for ROUGE-1, BERTScore, SummaC, and FactCC; for BARTScore, less negative values indicate better generation quality.

Dataset	**Model**	**R-1**		**BERT**		**BART**		**SummaC**		**FactCC**	
		Bef.	**Aft.**	**Bef.**	**Aft.**	**Bef.**	**Aft.**	**Bef.**	**Aft.**	**Bef.**	**Aft.**
CNN/DM	BART-Large	0.414	0.415	0.889	0.891	-2.402	-2.400	0.733	0.734	0.897	0.896
	T5-Small	0.432	0.433	0.892	0.894	-2.388	-2.386	0.737	0.738	0.901	0.900
	Pegasus-Large	0.421	0.420	0.890	0.891	-2.396	-2.395	0.734	0.735	0.898	0.897
	Flan-T5-Large	0.445	0.446	0.895	0.897	-2.375	-2.373	0.741	0.742	0.904	0.903
	Qwen2.5B	0.438	0.439	0.893	0.895	-2.382	-2.380	0.738	0.739	0.902	0.901
	Qwen-3	0.441	0.442	0.894	0.896	-2.379	-2.377	0.739	0.740	0.903	0.902
	Mistral-7B	0.436	0.435	0.891	0.893	-2.390	-2.388	0.736	0.737	0.900	0.899
	Vicuna-7B	0.433	0.434	0.890	0.892	-2.394	-2.392	0.735	0.736	0.899	0.898
	LLaMA-3-8B	0.448	0.449	0.896	0.898	-2.370	-2.368	0.742	0.743	0.905	0.904
	Gemma-3-it	0.451	0.452	0.897	0.899	-2.367	-2.365	0.743	0.744	0.906	0.905
	GPT-oss-20b	0.454	0.455	0.899	0.900	-2.362	-2.360	0.745	0.746	0.907	0.906
WikiSum	BART-Large	0.298	0.299	0.872	0.874	-2.455	-2.453	0.705	0.706	0.874	0.873
	T5-Small	0.315	0.317	0.876	0.878	-2.438	-2.436	0.711	0.712	0.879	0.878
	Pegasus-Large	0.305	0.304	0.874	0.875	-2.448	-2.447	0.708	0.709	0.876	0.875
	Flan-T5-Large	0.322	0.324	0.880	0.882	-2.428	-2.426	0.715	0.716	0.883	0.882
	Qwen2.5B	0.318	0.319	0.877	0.879	-2.434	-2.432	0.712	0.713	0.880	0.879
	Qwen-3	0.320	0.322	0.879	0.881	-2.431	-2.429	0.714	0.715	0.882	0.881
	Mistral-7B	0.316	0.317	0.876	0.878	-2.440	-2.438	0.711	0.712	0.879	0.878
	Vicuna-7B	0.314	0.315	0.875	0.877	-2.443	-2.441	0.710	0.711	0.878	0.877
	LLaMA-3-8B	0.323	0.324	0.881	0.883	-2.425	-2.423	0.716	0.717	0.884	0.883
	Gemma-3-it	0.326	0.327	0.882	0.884	-2.421	-2.419	0.717	0.718	0.885	0.884
	GPT-oss-20b	0.329	0.330	0.884	0.885	-2.417	-2.415	0.719	0.720	0.886	0.885
ArXiv	BART-Large	0.346	0.344	0.879	0.880	-2.430	-2.429	0.720	0.721	0.884	0.883
	T5-Small	0.358	0.359	0.883	0.885	-2.418	-2.416	0.724	0.725	0.889	0.888
	Pegasus-Large	0.352	0.351	0.881	0.882	-2.424	-2.423	0.722	0.723	0.886	0.885
	Flan-T5-Large	0.364	0.365	0.886	0.888	-2.407	-2.405	0.728	0.729	0.892	0.891
	Qwen2.5B	0.361	0.362	0.884	0.886	-2.412	-2.410	0.726	0.727	0.890	0.889
	Qwen-3	0.363	0.364	0.885	0.887	-2.409	-2.407	0.727	0.728	0.891	0.890
	Mistral-7B	0.360	0.361	0.883	0.885	-2.416	-2.414	0.724	0.725	0.889	0.888
	Vicuna-7B	0.358	0.359	0.882	0.884	-2.419	-2.417	0.723	0.724	0.888	0.887
	LLaMA-3-8B	0.366	0.367	0.887	0.889	-2.403	-2.401	0.729	0.730	0.893	0.892
	Gemma-3-it	0.369	0.370	0.888	0.890	-2.399	-2.397	0.730	0.731	0.894	0.893
	GPT-oss-20b	0.372	0.373	0.890	0.891	-2.395	-2.393	0.732	0.733	0.895	0.894
PubMed	BART-Large	0.372	0.371	0.884	0.885	-2.414	-2.413	0.727	0.728	0.890	0.889
	T5-Small	0.384	0.385	0.888	0.890	-2.402	-2.400	0.731	0.732	0.895	0.894
	Pegasus-Large	0.378	0.377	0.886	0.887	-2.408	-2.407	0.729	0.730	0.892	0.891
	Flan-T5-Large	0.389	0.390	0.891	0.893	-2.392	-2.390	0.735	0.736	0.898	0.897
	Qwen2.5B	0.386	0.387	0.889	0.891	-2.397	-2.395	0.733	0.734	0.896	0.895
	Qwen-3	0.388	0.389	0.890	0.892	-2.394	-2.392	0.734	0.735	0.897	0.896
	Mistral-7B	0.385	0.384	0.888	0.890	-2.401	-2.399	0.731	0.732	0.895	0.894
	Vicuna-7B	0.383	0.384	0.887	0.889	-2.404	-2.402	0.730	0.731	0.894	0.893
	LLaMA-3-8B	0.392	0.393	0.892	0.894	-2.388	-2.386	0.736	0.737	0.899	0.898
	Gemma-3-it	0.395	0.396	0.893	0.895	-2.384	-2.382	0.737	0.738	0.900	0.899
	GPT-oss-20b	0.398	0.399	0.895	0.896	-2.380	-2.378	0.739	0.740	0.901	0.900
Multi-XScience	BART-Large	0.331	0.330	0.875	0.877	-2.448	-2.446	0.707	0.708	0.878	0.877
	T5-Small	0.346	0.347	0.879	0.881	-2.432	-2.430	0.711	0.712	0.882	0.881
	Pegasus-Large	0.338	0.337	0.877	0.878	-2.440	-2.439	0.709	0.710	0.880	0.879
	Flan-T5-Large	0.351	0.352	0.883	0.885	-2.421	-2.419	0.715	0.716	0.886	0.885
	Qwen2.5B	0.348	0.349	0.881	0.883	-2.427	-2.425	0.713	0.714	0.884	0.883
	Qwen-3	0.350	0.351	0.882	0.884	-2.424	-2.422	0.714	0.715	0.885	0.884
	Mistral-7B	0.347	0.346	0.879	0.881	-2.431	-2.429	0.711	0.712	0.882	0.881
	Vicuna-7B	0.345	0.346	0.878	0.880	-2.435	-2.433	0.710	0.711	0.881	0.880
	LLaMA-3-8B	0.353	0.354	0.884	0.886	-2.416	-2.414	0.716	0.717	0.887	0.886
	Gemma-3-it	0.356	0.357	0.885	0.887	-2.412	-2.410	0.717	0.718	0.888	0.887
	GPT-oss-20b	0.359	0.360	0.887	0.888	-2.408	-2.406	0.719	0.720	0.889	0.888

Table 10: Comparison with lead-bias mitigation baselines on the remaining datasets using Lead Bias Ratio (LBR). Lower values indicate reduced positional shortcut reliance. MS: multi-stage training; Adv: AdvSumm.

Model	CNN/DM					WikiSum					ArXiv					PubMed				
	Orig	MS	Adv	NeuS	TVN	Orig	MS	Adv	NeuS	TVN	Orig	MS	Adv	NeuS	TVN	Orig	MS	Adv	NeuS	TVN
BART-Large	0.84	0.78	0.76	0.73	0.57	0.86	0.80	0.78	0.75	0.60	0.85	0.79	0.77	0.74	0.59	0.83	0.77	0.75	0.72	0.56
T5-Small	0.80	0.75	0.73	0.71	0.59	0.83	0.77	0.75	0.73	0.62	0.81	0.76	0.74	0.72	0.60	0.79	0.74	0.72	0.70	0.58
Pegasus-Large	0.81	0.76	0.74	0.72	0.55	0.84	0.78	0.76	0.74	0.58	0.82	0.77	0.75	0.73	0.57	0.80	0.75	0.73	0.71	0.54
Flan-T5-Large	0.79	0.76	0.74	0.72	0.42	0.82	0.78	0.76	0.74	0.45	0.80	0.77	0.75	0.73	0.43	0.78	0.75	0.73	0.71	0.40
Qwen2.5B	0.77	0.75	0.73	0.72	0.44	0.80	0.77	0.75	0.73	0.47	0.78	0.76	0.74	0.72	0.45	0.76	0.74	0.72	0.70	0.42
Qwen-3	0.76	0.74	0.72	0.71	0.42	0.79	0.76	0.74	0.72	0.45	0.77	0.75	0.73	0.71	0.43	0.75	0.73	0.71	0.69	0.40
Mistral-7B	0.78	0.76	0.74	0.72	0.45	0.81	0.78	0.76	0.74	0.48	0.79	0.77	0.75	0.73	0.46	0.77	0.75	0.73	0.71	0.43
Vicuna-7B	0.79	0.77	0.75	0.73	0.46	0.82	0.79	0.77	0.75	0.49	0.80	0.78	0.76	0.74	0.47	0.78	0.76	0.74	0.72	0.44
LLaMA-3-8B	0.77	0.75	0.73	0.71	0.43	0.80	0.77	0.75	0.73	0.46	0.78	0.76	0.74	0.72	0.44	0.76	0.74	0.72	0.70	0.41
Gemma-3-it	0.78	0.76	0.74	0.72	0.41	0.81	0.78	0.76	0.74	0.44	0.79	0.77	0.75	0.73	0.42	0.77	0.75	0.73	0.71	0.39
GPT-oss-20b	0.79	0.77	0.75	0.73	0.40	0.82	0.79	0.77	0.75	0.43	0.80	0.78	0.76	0.74	0.41	0.78	0.76	0.74	0.72	0.38

Table 11: Comparison with generic unlearning strategies on the remaining datasets using Lead Bias Ratio (LBR). Lower values indicate reduced positional shortcut reliance.

Model	CNN/DM				WikiSum				ArXiv				PubMed			
	Orig	Approx	Exact	TVN	Orig	Approx	Exact	TVN	Orig	Approx	Exact	TVN	Orig	Approx	Exact	TVN
BART-Large	0.84	0.75	0.72	0.57	0.86	0.77	0.74	0.60	0.85	0.76	0.73	0.59	0.83	0.74	0.71	0.56
T5-Small	0.80	0.72	0.69	0.59	0.83	0.75	0.72	0.62	0.81	0.73	0.70	0.60	0.79	0.71	0.68	0.58
Pegasus-Large	0.81	0.74	0.71	0.55	0.84	0.76	0.73	0.58	0.82	0.75	0.72	0.57	0.80	0.73	0.70	0.54
Flan-T5-Large	0.79	0.71	0.68	0.42	0.82	0.74	0.71	0.45	0.80	0.72	0.69	0.43	0.78	0.70	0.67	0.40
Qwen2.5B	0.77	0.70	0.67	0.44	0.80	0.72	0.69	0.47	0.78	0.71	0.68	0.45	0.76	0.69	0.66	0.42
Qwen-3	0.76	0.69	0.66	0.42	0.79	0.71	0.68	0.45	0.77	0.70	0.67	0.43	0.75	0.68	0.65	0.40
Mistral-7B	0.78	0.71	0.68	0.45	0.81	0.73	0.70	0.48	0.79	0.72	0.69	0.46	0.77	0.70	0.67	0.43
Vicuna-7B	0.79	0.72	0.69	0.46	0.82	0.74	0.71	0.49	0.80	0.73	0.70	0.47	0.78	0.71	0.68	0.44
LLaMA-3-8B	0.77	0.70	0.67	0.43	0.80	0.72	0.69	0.46	0.78	0.71	0.68	0.44	0.76	0.69	0.66	0.41
Gemma-3-it	0.78	0.70	0.66	0.41	0.81	0.72	0.68	0.44	0.79	0.71	0.67	0.42	0.77	0.69	0.65	0.39
GPT-oss-20b	0.79	0.71	0.67	0.40	0.82	0.73	0.69	0.43	0.80	0.72	0.68	0.41	0.78	0.70	0.66	0.38

Table 12: Comparison with robustness-oriented defenses on the remaining datasets using Lead Bias Ratio (LBR). Lower values indicate reduced positional shortcut reliance. Purif.: text adversarial purification; Lang.: Language-Tool; AdvAug: adversarial data augmentation.

Model	CNN/DM					WikiSum					ArXiv					PubMed				
	Purif.	scRNN	Lang.	AdvAug	TVN	Purif.	scRNN	Lang.	AdvAug	TVN	Purif.	scRNN	Lang.	AdvAug	TVN	Purif.	scRNN	Lang.	AdvAug	TVN
BART-Large	0.80	0.82	0.81	0.70	0.57	0.82	0.84	0.83	0.72	0.60	0.81	0.83	0.82	0.71	0.59	0.79	0.81	0.80	0.69	0.56
T5-Small	0.76	0.77	0.76	0.67	0.59	0.79	0.80	0.79	0.70	0.62	0.77	0.78	0.77	0.68	0.60	0.75	0.76	0.75	0.66	0.58
Pegasus-Large	0.76	0.77	0.77	0.67	0.55	0.80	0.81	0.80	0.70	0.58	0.78	0.79	0.78	0.68	0.57	0.76	0.77	0.76	0.66	0.54
Flan-T5-Large	0.74	0.75	0.74	0.64	0.42	0.77	0.78	0.77	0.67	0.45	0.75	0.76	0.75	0.65	0.43	0.73	0.74	0.73	0.63	0.40
Qwen2.5B	0.72	0.73	0.72	0.62	0.44	0.75	0.76	0.75	0.65	0.47	0.73	0.74	0.73	0.63	0.45	0.71	0.72	0.71	0.61	0.42
Qwen-3	0.71	0.72	0.71	0.61	0.42	0.74	0.75	0.74	0.64	0.45	0.72	0.73	0.72	0.62	0.43	0.70	0.71	0.70	0.60	0.40
Mistral-7B	0.73	0.74	0.73	0.63	0.45	0.76	0.77	0.76	0.66	0.48	0.74	0.75	0.74	0.64	0.46	0.72	0.73	0.72	0.62	0.43
Vicuna-7B	0.74	0.75	0.74	0.64	0.46	0.77	0.78	0.77	0.67	0.49	0.75	0.76	0.75	0.65	0.47	0.73	0.74	0.73	0.63	0.44
LLaMA-3-8B	0.72	0.73	0.72	0.62	0.43	0.75	0.76	0.75	0.65	0.46	0.73	0.74	0.73	0.63	0.44	0.71	0.72	0.71	0.61	0.41
Gemma-3-it	0.72	0.73	0.72	0.61	0.41	0.76	0.77	0.76	0.65	0.44	0.74	0.75	0.74	0.63	0.42	0.72	0.73	0.72	0.61	0.39
GPT-oss-20b	0.73	0.74	0.73	0.61	0.40	0.77	0.78	0.77	0.65	0.43	0.75	0.76	0.75	0.63	0.41	0.73	0.74	0.73	0.61	0.38