Less Is More? Examining Fairness in Pruned Large Language Models for Summarising Opinions

Nannan Huang
RMIT University, Australia
amber.huang@student.rmit.edu.au

Haytham M. Fayek
RMIT University, Australia
haytham.fayek@ieee.org

Xiuzhen Zhang
RMIT University, Australia
xiuzhen.zhang@rmit.edu.au

Abstract

Model compression through post-training pruning offers a way to reduce model size and computational requirements without significantly impacting model performance. However, the effect of pruning on the fairness of LLM-generated summaries remains unexplored, particularly for opinion summarisation where biased outputs could influence public views. In this paper, we present a comprehensive empirical analysis of opinion summarisation, examining three state-of-the-art pruning methods and various calibration sets across three open-source LLMs using four fairness metrics. Our systematic analysis reveals that pruning methods have a greater impact on fairness than calibration sets. Building on these insights, we propose High Gradient Low Activation (HGLA) pruning, which identifies and removes parameters that are redundant for input processing but influential in output generation. Our experiments demonstrate that HGLA can better maintain or even improve fairness compared to existing methods, showing promise across models and tasks where traditional methods have limitations. Our human evaluation shows HGLA-generated outputs are fairer than existing state-of-the-art pruning methods. Code is available at: https://github.com/amberhuang01/HGLA.

1 Introduction

Scaling language models has led to emergent capabilities (Brown, 2020; Radford et al., 2019; Chowdhery et al., 2023; Touvron et al., 2023; Le Scao et al., 2023), however, it has also led to increased computational resources (Wu et al., 2022; Zhu et al., 2023; Menghani, 2023). While reducing the size of LLMs through post-training pruning has become popular and attracted substantial research interest (Frantar and Alistarh, 2023; Sun et al., 2023; Das et al., 2023), their impact beyond performance metrics remains largely understudied, with the work by Chrysostomou et al. (2024) on model hallucination being one of the only studies that investigates pruning from a non-performance angle.

Another critical but unexplored aspect of pruning is its impact on model fairness. The use of AI language systems with inherent biases could shape the way audiences interpret and process information (Jakesch et al., 2023; Durmus et al., 2023; Epstein et al., 2023), especially in tasks such as opinion summarisation where biased outputs could significantly influence public opinion. It is well-acknowledged that LLMs were exposed to uncurated data that may contain societal bias, which inevitably perpetuates social stereotypes in models (Vig et al., 2020; Sheng et al., 2019; Liang et al., 2021; Gallegos et al., 2024; Li et al., 2023) and propagates to downstream tasks (Feng et al., 2023).

Post-training pruning typically uses a small calibration set of samples to identify parameters for removal. Given the fairness concerns, it is crucial to examine all aspects of post-training pruning procedures that could impact model fairness, including both pruning methods and the calibration set. While recent work has shown that the choice of calibration set significantly impacts model performance (Williams and Aletras, 2024), its effect on model fairness remains unexplored. This gap is particularly concerning given the established relationship between training data and model bias.

To the best of our knowledge, our work presents the first systematic investigation of how pruning affects fairness in LLM-generated summaries. Through a comprehensive analysis using three state-of-the-art pruning methods, various calibration sets across three large language models, and four fairness metrics, our results demonstrate that pruning methods have a greater influence on model fairness than calibration sets. We find that pruning more model parameters can lead to decreased fairness while maintaining performance, suggesting

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17994–18018
November 4-9, 2025 ©2025 Association for Computational Linguistics

that performance-focused pruning methods may inadvertently amplify biases. This complex relationship varies across tasks, pruning methods, and calibration sets, highlighting the critical importance of carefully selecting pruning methods and continuously monitoring fairness metrics during the pruning process.

In response to these limitations, we introduce High Gradient Low Activation (HGLA) pruning using calibration sets—a novel procedure that identifies and removes parameters that are redundant in input processing (i.e., low activation) but influential in output generation (i.e., high gradient), offering a promising method to maintain or even improve model fairness during the post-training pruning process. Our human evaluation study demonstrates that this pruning procedure produces fairer summaries compared to other state-of-the-art post-pruning methods.

2 Related Work

2.1 Post-training Pruning for LLMs

Model pruning aims to reduce model size while maintaining performance. Magnitude pruning examines the absolute value of model weights as a simple baseline. The activation magnitudes indicate how parameters contribute to processing inputs. By analysing activation strengths using calibration data, we can identify less crucial parameters, such as those showing consistently weak activations that may be unnecessary and redundant in processing input. Recent post-training pruning methods such as SparseGPT (Frantar and Alistarh, 2023) and Wanda (Sun et al., 2023) are mainly guided using activation information to prune such redundant parameters. A model's gradients provide information on how sensitive its output is with respect to its parameters. Parameters with larger gradients are more sensitive on the model's output, as small changes to these parameters would significantly change the output. GBLM-Pruner (Das et al., 2023) combines the information in both activation and gradient with a scaling factor to decide how much contribution is made by the gradient. Parameters that produce the lowest gradient and activation should be removed since removing these parameters has the minimum impact on both the output of the model and also how the model processes input information. These methods require calibration data, typically sampled from C4 (Raffel et al., 2020). While research shows calibration set selection significantly impacts model performance (Williams and Aletras, 2024), its impact on model fairness has not been studied.

While LLMs can maintain performance with up to 50% pruning (Jaiswal et al., 2023), pruning's effect on fairness remains unexplored. Prior work has shown pruning can reduce hallucination by increasing source document reliance (Chrysostomou et al., 2024) but the impact on model fairness has not been studied. The key contributions of this work are: (1) evaluating fairness across multiple pruning methods, (2) examining effects of different calibration sets, and (3) based on the limitations identified in existing methods, we propose a novel pruning procedure that addresses these challenges.

2.2 Fairness in Summarising Opinions

Prior work has explored social bias in language models across gender, race, and other attributes (Vig et al., 2020; Sheng et al., 2019; Liang et al., 2021; Ladhak et al., 2023; Santurkar et al., 2023), showing how these biases propagate from training data to downstream tasks (Feng et al., 2023). In opinion summarisation, fair models should represent diverse opinions proportionally to their source documents (Shandilya et al., 2018), and biased outputs risk misrepresenting input distributions. Prior research has examined fairness across demographic attributes such as gender, race, political orientation (Dash et al., 2019), dialect (Blodgett et al., 2016), and opinion diversity (Huang et al., 2023).

Recent studies have investigated fairness in fine-tuning (Huang et al., 2024) and prompt-based models (Zhang et al., 2023). However, the impact of post-training pruning on fairness in abstractive summarisation remains unexplored.

3 High Gradient Low Activation Pruning (HGLA)

Formally, pruning methods assign a significance score $\mathbf{S}_{i,j}$ to each element of a layer's weight matrix $\mathbf{W}_{i,j}$. State-of-the-art post-training pruning methods can also factor in additional information, such as the layer's input activations $\mathbf{X}$ or gradients $\mathbf{G}$, derived from a calibration dataset.

While traditional pruning methods remove parameters based on either low activation or low gradient values to maintain model performance, our fairness-aware approach uses a calibration set to identify a specific type of parameters: those showing low activation but high gradients. These pa-

rameters, while not actively processing input features (i.e., low activation), significantly influence the model's output (i.e., high gradients). By targeting such parameters using examples from the calibration set, we aim to modify the model's behaviour to generate less biased outputs that diverge from the vanilla model's outputs.

The procedure focuses on using a calibration set to isolate and identify parameters that are less activated when processing information from a specific perspective (i.e., redundant in processing information representing a specific side), while removing these weights would produce different outputs (i.e., generate less biased outputs that diverge from the vanilla model's outputs). Our hypothesis is that these parameters may encode patterns that affect output generation while being non-essential for input processing. By identifying and removing such parameters, we aim to maintain or potentially enhance model fairness while preserving the model's ability to effectively process input information.

Given a weight matrix, $\mathbf{W}$, a gradient matrix, $\mathbf{G}$, and an input feature activation, $\mathbf{X}$, the goal is to generate the masking matrix $\mathbf{W}_m$. We followed the work by Sun et al. (2023) in using the ℓ_2 norm in measuring activation magnitudes and Das et al. (2023) in using the ℓ_2 normalisation across samples' gradients as follows:

$$\mathbf{W}_{\mathrm{m}} = \left| \frac{|\mathbf{W}[i,j]| \cdot \|\mathbf{X}[:,j]\|_2}{|\mathbf{G}[:,i,j]|_p} \right| \tag{1}$$

In our preliminary analysis, we found that Equation 1 is dominated by the inverse of gradients with high magnitude, which resulted in removing parameters with these gradients, that led in some instances to model collapse. We therefore normalised the inverse of the gradients so that the magnitude of the activations and gradients will be on the same scale as follows:

$$|\mathbf{G}'[:,i,j]|_p = |\mathbf{G}[:,i,j]|_p \cdot \frac{\frac{1}{mn}\sum_{i,j}(|\mathbf{W}[i,j]| \cdot \|\mathbf{X}[:,j]\|_2)}{\frac{1}{mn}\sum_{i,j}|\mathbf{G}[:,i,j]|_p} \tag{2}$$

4 Experiments

4.1 Datasets

We include the following datasets in our study: (1) MOS (Bilal et al., 2022) (2) FewSum (Bražinskas et al., 2022) (3) AmaSum (Bražinskas et al., 2021) (4) FairSumm (Dash et al., 2019) (5) Amazon reviews 2023 (Hou et al., 2024). [1] Our experimental framework utilises three functionally distinct dataset categories: performance evaluation, fairness evaluation and calibration datasets.

Performance evaluation datasets quantitatively assess summarisation quality preservation using established automatic performance evaluation metrics mentioned in Section 4.5 against gold-standard reference summaries. For political tweet summarisation, we employ the manually validated test set from the political partition of MOS (Bilal et al., 2022), which contains coherent opinion collections for evaluating summarisation performance on political discourse. For product review summarisation, we utilise the established test sets from FewSum (Bražinskas et al., 2022) and AmaSum (Bražinskas et al., 2021), which provide human-annotated reference summaries for comprehensive quality assessment.

Fairness evaluation datasets measure opinion representation balance through fairness metrics mentioned in Section 4.5. We construct separate test sets specifically designed to measure bias, built from FairSumm (Dash et al., 2019) for political tweet fairness evaluation and Amazon Reviews 2023 (Hou et al., 2024) for product review fairness evaluation. These fairness evaluation datasets are structurally independent from the performance evaluation test sets described above, as fairness assessment requires balanced opinion distributions rather than reference summaries. Complete fairness dataset construction methodology is documented in Appendix A.2.

Calibration datasets inform parameter selection during the pruning process and are constructed from the aforementioned source datasets using methodology detailed in Section 4.4. These serve exclusively as algorithmic guidance for pruning rather than evaluation benchmarks.

4.2 Models

We use three popular open-source LLMs: (1) TinyLlama (Zhang et al., 2024) (2) Gemma (Team et al., 2024) (3) Llama 3 (Dubey et al., 2024). We use both base models (Llama 3 and TinyLlama) as well as instruction-tuned models (Gemma). We select these models to cover a diverse range of model sizes ranging from 1.1B, 2B and 8B. The prompt template we use is in Section A.3.

[1] https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

4.3 Baseline Pruning Methods

We compare different SOTA post-training pruning methods and their effects on fairness, including: (1) Magnitude (Han et al., 2015), (2) SparseGPT (Jakesch et al., 2023), (3) Wanda (Sun et al., 2023), and (4) GBLM-Pruner (Das et al., 2023).

We perform unstructured pruning for all pruning methods since it offers finer control over weight retention, allowing us to preserve crucial features that may be important for fair representation of minority groups. For magnitude, SparseGPT and Wanda we use the implementation provided by Sun et al. (2023). [2] For GBLM-Pruner, we use the implementation provided by Das et al. (2023), [3] we use both the gradient only version denoted as GBLM-Gradient and gradient with activation version denoted as GBLM-Pruner for the remaining of the paper. We use the ℓ_2 norm of the gradients and the scaling factor α, we use a value of 100 as suggested in GBLM-Pruner (Das et al., 2023) to account for the small magnitude of gradients when combining activations and gradients.

4.4 Calibration Sets

Each calibration set consists of 128 input collections, with each collection containing either 30 political tweets from FairSumm (Dash et al., 2019) or 8 reviews from Amazon Reviews 2023 (Hou et al., 2024). From the *input perspective*, we directly utilise labels to construct calibration sets. From the *output perspective*, we randomly construct 50,000 input collections and use the vanilla model to generate summaries. We then construct the calibration set by sampling based on the output SPD according to specific conditions. The descriptions of each calibration set we construct are as follows:

- **Single-sided input** contains only single-sided or biased information by selecting input from the one only until the target length is reached.
- **Fair input** contains information from both sides in equal proportion (i.e., selecting 8 reviews for the same product: 4 with positive opinions and 4 with negative opinions).
- **Mixed input** includes a balanced mix of fair and biased inputs, with 128 total input collections. Half of these contain single-sided information, while the other half contain balanced information from both perspectives.
- **Biased output** based on the output SPD, we construct the calibration set by selecting inputs that produce extreme SPD values (i.e., positive or negative 1) using the vanilla model.
- **Fair output** based on the output SPD, we construct the calibration set by selecting inputs that produce fair SPD values (i.e., SPD is 0) using the vanilla model.
- **Mixed output** based on the output SPD, we create a mixed calibration set that includes inputs producing both fair and biased outputs. Half of the calibration set produces biased output, while the remaining half produces fair output, as determined using the vanilla model.

4.5 Evaluation Metrics

Performance: the quality of the generated summaries is evaluated through comparison with the corresponding reference summaries using two automatic evaluation metrics, ROUGE (Lin, 2004), an n-gram overlapping metric, and, BERTScore (Zhang et al., 2019), an embedding-based evaluation metric. For BERTScore, we use the F1 measure. For ROUGE, we use ROUGE-1 and ROUGE-2 for unigrams and bigrams, alongside ROUGE-L, which measures the longest common overlapped sequence between generated and reference summaries.

Fairness: A fair model should represent diverse groups equally or proportionally to their population. We evaluate model fairness using multiple metrics: Second-Order SPD (SPD) (Huang et al., 2024), Binary Unfair Rate (BUR), Unfair Error Rate (UER), and Second-Order Fairness (SOF) (Zhang et al., 2023).

- **Second-Order SPD (SPD)** (Huang et al., 2024): Evaluates sentence-level social attributes using a fine-tuned model, comparing distributions between summaries and source documents.
- **Binary Unfair Rate (BUR)** (Zhang et al., 2023): Measures overall fairness by calculating the ratio of fair summaries to the total number of generated summaries.
- **Unfair Error Rate (UER)**: Evaluates underrepresentation in the generated summaries by calculating the average difference between target and generated social value distributions.

[2] https://github.com/locuslab/wanda
[3] https://github.com/VILA-Lab/GBLM-Pruner

- **Second-Order Fairness (SOF)**: Assesses the spread of unfairness across different values by calculating the variance of UER across all values, which highlights which values are subjected to more unfairness in each sample.

To calculate SPD, we classify sentences in the generated summaries using steps and models described in Section A.1, then compare proportions of different opinions. For BUR, UER, and SOF, we follow the methodology proposed by Zhang et al. (2023), using the average of n-gram, BERTScore (Zhang et al., 2019), and BARTScore (Yuan et al., 2021) matching. We use the publicly available implementation. [4]

4.6 Implementation Details

We use the model implementation and weights available from Hugging Face (Wolf et al., 2020). We perform experiments using either one or two NVIDIA A100 (40GB) GPUs. For the pruning methods, we use the hyperparameters provided by Frantar and Alistarh (2023), Sun et al. (2023) and Das et al. (2023). For calibration sets, we make a slight modification, instead of using an input length of 2048 tokens, we use 512 since we observed that most of the input lengths in summarising political tweets and reviews are around 512 rather than 2048.

5 Results and Discussion

5.1 Pruning and Summarisation Performance

We evaluate model performance using ROUGE 1 (Lin, 2004) and BERTScore (Zhang et al., 2019) with random calibration sets (construction details in Section A.4). Results are visualised in Figure 1, with full details in Appendix A.5 and performance-fairness tradeoffs in Figure 3. Magnitude pruning (Han et al., 2015) shows significant degradation compared to other methods, while SparseGPT (Jakesch et al., 2023) maintains performance best at 50% sparsity, except for Gemma-2B (Team et al., 2024) on review summarisation. Given performance degradation at 50% sparsity across most methods, we eliminate magnitude pruning and cap sparsity at 40% for fairness comparisons.

5.2 Pruning, Calibration Set, and Fairness

Previous studies found that calibration sets used in pruning methods play an important role in model performance (Williams and Aletras, 2024). Therefore, to examine model fairness, we use the curated calibration sets mentioned in Section 4.4.

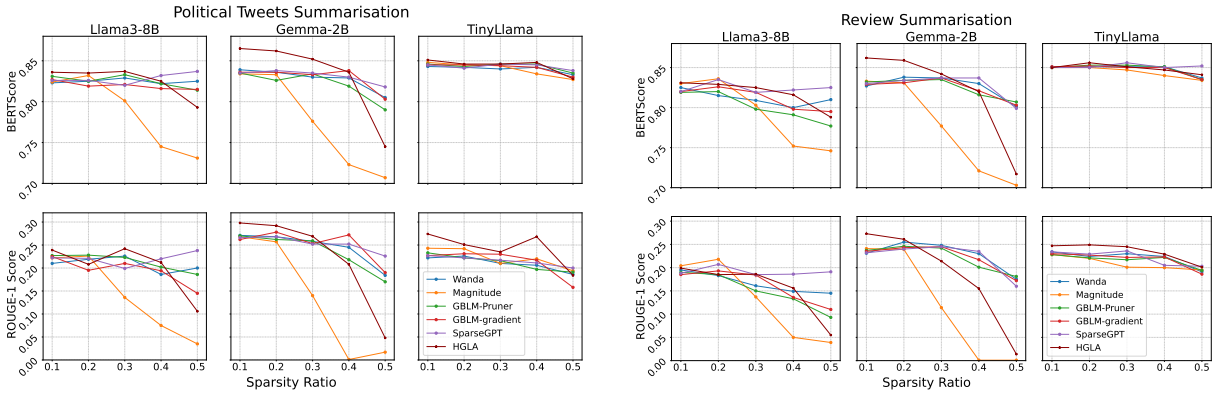
We evaluate fairness using multiple metrics mentioned in Section 4.5 on both vanilla and pruned models, quantifying improvements through the absolute difference between their respective metrics. Improvements are considered positive when this difference falls between zero and the vanilla model's metric value. To provide a comprehensive analysis and visualise the complex interplay between pruning methods, calibration sets, and pruning ratios, we present results in Figure 2 in two ways: averaging across different pruning methods for each calibration set, and averaging across calibration sets for each pruning method. Detailed model performance and fairness results are provided in Appendix A.6 and Appendix A.7.

From the calibration set perspective, our analysis reveals interesting patterns in fairness improvements. When examining the average improvement across methods per calibration set (shown in the first row of each sub-graph in Figure 2), we observe that the trend lines from different calibration sets (shown in the second row of each sub-graph in Figure 2) exhibit frequent intersections and maintain close proximity to each other. The variations across calibration sets are small, while differences across pruning methods are more pronounced, indicating that pruning methods have a greater impact on model fairness. This observation is quantitatively supported by variance analysis detailed in Appendix A.9.

From the pruning method perspective, increasing sparsity ratios generally leads to decreased model fairness across different metrics in both political tweet and review summarisation. Our analysis demonstrates that HGLA consistently outperforms or matches alternative methods for fairness improvement across different summarisation tasks and calibration sets (evidenced by its position in the top two performance lines across metrics and summarisation datasets). Wanda also demonstrates competitive performance as a pruning method, exhibiting stable results across different sparsity ratios compared to other approaches. It maintains better fairness scores or shows less degradation as sparsity increases.

Overall, the relationship between model pruning and fairness shows general trends towards decreased fairness with increased pruning across summarisation tasks and metrics. The magnitude of this

[4] https://github.com/psunlpgroup/FairSumm



(a) Political tweet summarisation

(b) Review summarisation

Figure 1: BERTScore and ROUGE-1 for different pruning methods across political tweet summarisation and review summarisation. Apart from TinyLlama, models using magnitude pruning show significant degradation after a 20% sparsity ratio. SparseGPT generally offers the most robust model performance.

decline varies across tasks, methods, and evaluation metrics, with HGLA demonstrating enhanced fairness preservation and Wanda demonstrating more stable performance compared to other pruning methods. Suggesting these factors interact in complex and sometimes counterintuitive ways, with certain combinations of pruning methods and tasks showing unexpected preservation or degradation of fairness metrics. The results indicate that pruning without awareness of fairness implications could potentially harm model fairness, highlighting the importance of careful pruning method selection and continuous monitoring of fairness metrics during the pruning process.

In contrast to previous study (Chrysostomou et al., 2024), which found that increased pruning reduced hallucination by making models rely more on their original input, our findings suggest a different pattern for opinionated text summarisation. We found that pruning more does not necessarily lead to less biased summaries of opinionated text. One possible explanation is that model compression can cause models to forget or perform worse on minority classes and edge cases, which may amplify existing biases or create new disparities across different social groups (Hooker et al., 2019, 2020; Misra et al., 2024). Without explicit guidance from fairness metrics during pruning, unconstrained pruning could potentially compromise model fairness (Zayed et al., 2024).

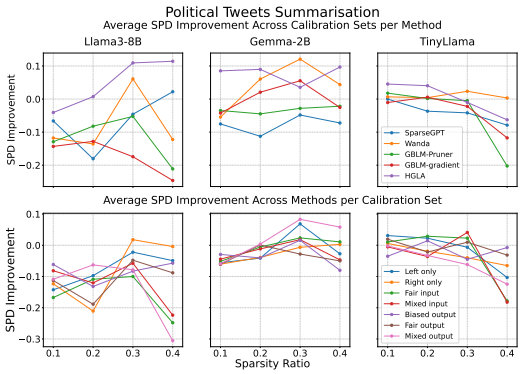
5.3 Fairness Pruned with HGLA

Existing SOTA post-training pruning methods that focus solely on model performance risk sacrificing fairness, as demonstrated by our findings in Section 5.2. This limitation becomes particularly crucial in opinion summarisation, where models can maintain performance while amplifying biases. Our method addresses this by targeting parameters that exhibit high-gradient and low-activation (HGLA) values, those that are redundant in processing input but sensitive to generation.

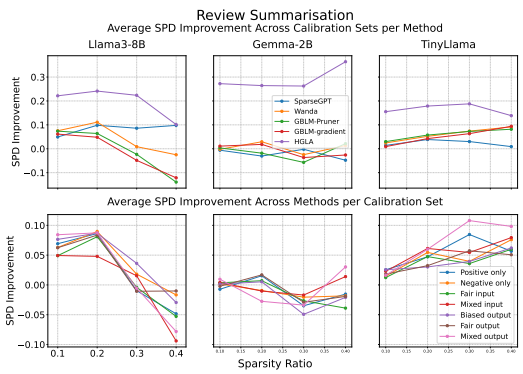
Section 5.2 presented aggregated results averaged across calibration sets to identify broad patterns in how pruning methods affect fairness. While this aggregation was necessary to identify general trends, it potentially obscures important nuances about specific calibration-pruning interactions. Table 1 offers a closer look at different calibration sets coupled with the preferred methods we found earlier—HGLA and Wanda (Sun et al., 2023). Specifically, we use single-sided information to isolate weights that are redundant in input processing of a particular side but whose changes would modify the output, allowing us to target parameters based on specific calibration conditions.

Our analysis reveals that HGLA demonstrates superior performance in fairness metrics compared to Wanda across tested models in both political and review summarisation. In political summarisation, HGLA shows particularly strong improvements, evidenced by greater improvements in both right-leaning and left-leaning information pruning overall. Specifically, it shows greater improvement in Gemma-2B and TinyLlama, while also enabling Llama3-8B to be pruned up to 40%.

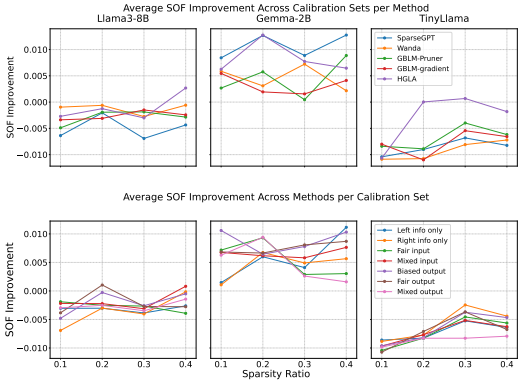
In review summarisation, while the improvements are more modest, HGLA maintains consistent performance across both positive and negative information pruning. In pruning using positive in-



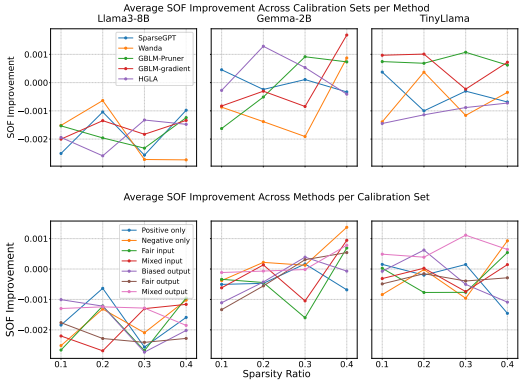
(a) Political tweet summarisation - SPD



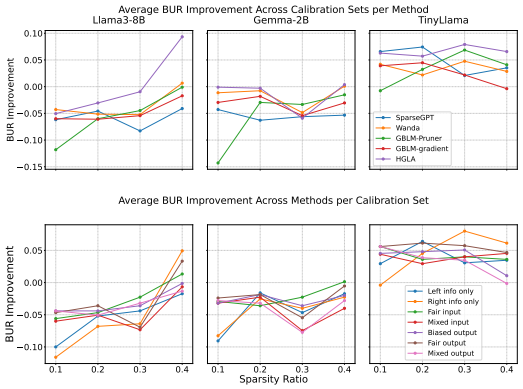
(b) Review summarisation - SPD



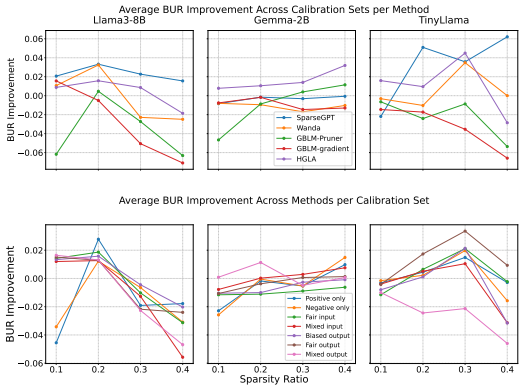
(c) Political tweet summarisation - SOF



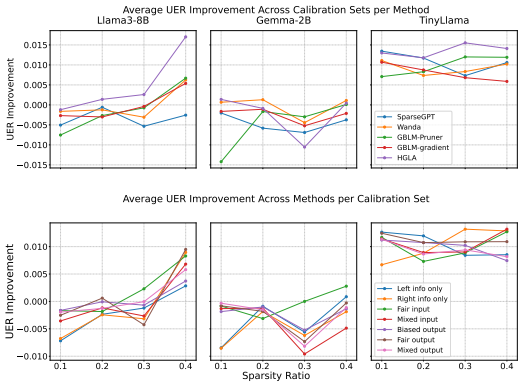
(d) Review summarisation - SOF



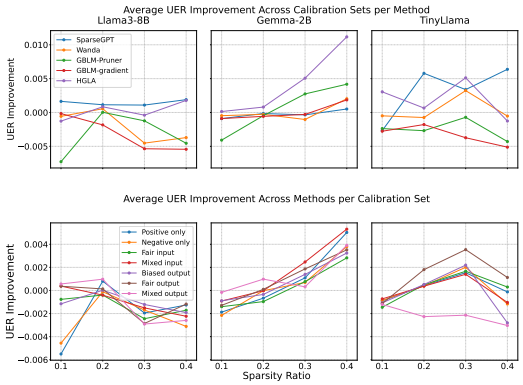
(e) Political tweet summarisation - BUR



(f) Review summarisation - BUR



(g) Political tweet summarisation - UER



(h) Review summarisation - UER

Figure 2: Impact on fairness from pruning methods and calibration sets. The figures in the first row show the average model fairness improvement across calibration sets for each pruning method. The figures in the second row display the average fairness improvement across pruning methods for each calibration set. We observe that the pruning method has a bigger impact on model fairness than calibration sets.

	Llama3-8B								Gemma-2B								TinyLlama							
	SPD		SOF		BUR		UER		SPD		SOF		BUR		UER		SPD		SOF		BUR		UER	
Ratio	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA
0.1	-0.106	-0.060	-0.005	-0.006	-0.113	-0.107	-0.005	0.006	-0.060	-0.046	0.005	0.005	0.000	0.013	0.000	0.002	-0.013	0.038	-0.012	-0.007	0.027	0.040	0.013	0.011
0.2	-0.142	-0.004	-0.001	-0.004	-0.053	-0.040	-0.004	0.000	0.043	0.045	0.007	0.010	0.007	-0.027	0.001	-0.002	-0.077	0.077	-0.004	0.003	0.013	0.040	0.006	0.010
0.3	0.082	0.119	-0.003	-0.005	-0.007	-0.060	-0.005	-0.001	0.225	-0.052	0.007	0.004	-0.040	-0.053	0.002	-0.013	0.006	-0.111	-0.010	0.004	0.013	0.147	0.008	0.024
0.4	-0.128	0.170	0.003	0.003	-0.020	0.100	0.008	0.015	0.038	0.112	-0.003	0.015	0.013	0.047	-0.006	0.004	0.006	-0.258	-0.003	0.010	0.020	0.067	0.015	0.011

(a) Political - Right information only pruned

	Llama3-8B								Gemma-2B								TinyLlama							
	SPD		SOF		BUR		UER		SPD		SOF		BUR		UER		SPD		SOF		BUR		UER	
Ratio	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA
0.1	-0.132	-0.232	0.001	-0.003	-0.100	-0.033	-0.008	-0.001	-0.014	-0.035	0.006	0.005	0.000	-0.007	0.001	0.001	0.006	0.065	-0.012	-0.012	0.060	0.073	0.009	0.015
0.2	-0.140	-0.134	-0.004	0.004	-0.060	-0.020	-0.002	0.002	0.039	-0.059	-0.002	0.015	0.020	0.007	0.003	0.000	-0.025	-0.037	-0.012	-0.003	0.040	0.073	0.005	0.015
0.3	0.183	0.043	-0.006	-0.002	-0.100	0.007	0.001	0.002	0.095	-0.205	0.007	0.009	0.000	-0.053	-0.004	-0.008	-0.015	0.037	-0.011	-0.007	0.073	0.067	0.006	0.010
0.4	-0.156	0.092	-0.001	0.001	0.040	0.100	0.003	0.019	-0.030	-0.153	0.004	0.017	-0.027	-0.007	0.005	0.002	-0.082	0.040	-0.009	-0.004	0.053	0.087	0.007	0.010

(b) Political - Left information only pruned

	Llama3-8B								Gemma-2B								TinyLlama							
	SPD		SOF		BUR		UER		SPD		SOF		BUR		UER		SPD		SOF		BUR		UER	
Ratio	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA
0.1	0.081	0.020	-0.004	0.000	0.000	-0.008	-0.001	-0.003	-0.001	-0.013	0.000	-0.001	0.021	0.020	0.002	0.001	0.016	0.031	-0.001	-0.001	-0.010	0.018	-0.001	0.003
0.2	0.130	0.097	0.002	-0.003	0.044	0.034	0.001	0.004	0.029	-0.028	-0.002	0.002	-0.016	0.016	-0.001	0.000	0.045	0.077	0.001	-0.001	-0.021	0.032	-0.002	0.002
0.3	-0.004	0.077	-0.002	-0.003	-0.039	0.016	-0.004	0.000	0.015	-0.032	-0.003	-0.001	-0.017	0.004	0.000	0.005	0.070	0.064	-0.001	-0.001	0.014	0.039	0.002	0.005
0.4	-0.045	-0.040	-0.003	0.001	-0.019	0.010	-0.004	0.003	0.008	0.140	-0.003	-0.003	0.008	0.051	0.005	0.012	0.126	0.061	-0.002	-0.001	0.010	-0.013	0.000	-0.001

(c) Review - Positive information only pruned

	Llama3-8B								Gemma-2B								TinyLlama							
	SPD		SOF		BUR		UER		SPD		SOF		BUR		UER		SPD		SOF		BUR		UER	
Ratio	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA	Wanda	HGLA
0.1	0.115	0.085	-0.003	-0.003	0.003	0.010	-0.001	-0.001	0.000	-0.013	0.001	0.001	0.020	0.003	0.001	0.000	0.040	0.029	-0.001	-0.002	0.003	0.022	0.000	0.003
0.2	0.102	0.145	0.001	-0.003	0.017	0.017	-0.001	0.000	0.033	-0.034	0.000	0.001	-0.017	0.007	-0.001	0.000	0.038	0.080	0.001	-0.001	-0.006	-0.004	-0.002	0.000
0.3	0.001	0.114	-0.003	0.001	-0.023	0.031	-0.004	0.001	-0.023	-0.004	0.000	-0.003	-0.007	0.021	-0.001	0.005	0.092	0.101	-0.002	-0.001	0.039	0.058	0.003	0.007
0.4	-0.050	-0.112	-0.003	-0.001	-0.013	-0.018	-0.002	0.001	0.027	0.219	0.002	-0.001	-0.016	0.033	0.001	0.007	0.072	-0.065	0.001	0.000	-0.018	-0.013	-0.002	0.000

(d) Review - Negative information only pruned

Table 1: Fairness improvement through pruning based on HGLA and Wanda, darker colours indicating greater fairness improvement. Compared to Wanda (Sun et al., 2023), HGLA brings greater benefits when using single-sided input opposite to the model's intrinsic bias, enabling deeper pruning (40% vs 30% for Llama3-8B) and showing improvements where Wanda produced minimal changes (e.g., Gemma-2B in review summarisation).

formation, HGLA achieves stable improvements, particularly in Gemma-2B and TinyLlama. Similarly, in pruning using negative information, HGLA demonstrates reliable fairness preservation, with notable improvements in metrics across all three models, suggesting its robust capability in maintaining fairness during pruning.

5.4 Human Evaluation

Rank	Model	Rating
1	HGLA	1450.6
2	SparseGPT	1420.3
3	Wanda	1404.2
4	GBLM-Pruner	1374.9
5	GBLM-Gradient	1350.0

Table 2: Elo ratings of pruning methods based on pairwise fairness comparisons. HGLA achieves the highest fairness rating, followed by activation-based methods such as SparseGPT and Wanda, and gradient-based GBLM variants in the lowest tier.

We conduct a systematic comparison of output fairness across five pruning methods using human evaluation. We use TinyLlama summaries with 0.3 sparsity ratio and negative-only review calibration, as this configuration demonstrated improved fairness. For our initial evaluation, we create 20 direct comparison pairs by randomly selecting outputs from different pruning methods. This approach allows us to verify human annotators' reliability. Following Shandilya et al. (2020), annotators identified distinct positive and negative opinions in each input, then assess which summary better preserves the original opinion distribution, achieving substantial inter-annotator agreement with a Fleiss' Kappa coefficient of 0.555 (Fleiss, 1971). Given the strong agreement, we extend the evaluation to 100 random comparison pairs to ensure a minimum of 30 comparisons per method, based on evidence that rating convergence occurs after approximately 30 comparative assessments (Ratings, 2024). Detailed information is provided in Appendix A.10.

We analyse the pairwise comparisons using the Elo rating system (Elo, 1967), a framework originally designed for chess rankings. The system dynamically adjusts ratings based on comparison outcomes and competitors' relative strength, with victories against higher-rated opponents earning more points. We initialise all methods with a default rating of 1400 and use a K-factor of 16, determining winners through majority voting across the three annotations per comparison.

As shown in Table 2, by aggregating human evaluations of 30 direct comparisons of summaries generated by each pruning method, the Elo rat-

ing system shows that HGLA ranks highest among all methods, followed by activation-based methods such as SparseGPT and Wanda, with gradient-based GBLM variants in the lowest tier. The emergence of three distinct performance tiers suggests that the selection of pruning method significantly influences the preservation of fairness characteristics in the resulting models.

6 Conclusion

In this study, we examined how post-training pruning affects LLM fairness in opinion summarisation, finding pruning methods impact fairness more than calibration set selection despite prior research emphasising calibration's importance. Unlike hallucination reduction from pruning, fairness decreases with increased pruning using SOTA methods. Not all pruning methods equally preserve model performance and fairness, highlighting the risk of overlooking fairness considerations when focusing solely on performance. To address these limitations, we introduce High Gradient Low Activation (HGLA) pruning, targeting parameters redundant in input processing but influential in output generation. Our examination shows promise for improving fairness across models and tasks where existing methods fall short. Human evaluation reveals that outputs generated by HGLA achieve better fairness compared to existing pruning methods. Our work emphasises that efficiency should not compromise fairness, and future research should explore the relationship between efficiency, performance, and fairness.

Limitations

We primarily concentrate on open-source language models due to the restricted access to parameters in closed-source LLMs. However, it's worth noting that our approach to assessing fairness in LLM-generated summaries of text with opinions remains relevant for researchers with access to closed-source models. Our research deliberately focuses on post-training pruning as it uniquely allows us to investigate how selectively removing specific parameters affects model bias—a question that other compression techniques cannot address in the same way. Unlike distillation or quantisation, which transform representations or reduce precision uniformly, pruning provides a distinctive analytical lens to observe how different regions of the parameter space influence biased behaviour. Our core research question examines whether it is possible to strategically remove parts of a model to reduce bias—a question fundamentally aligned with pruning's selective removal approach. Additionally, our work focus on unstructured pruning methods only since unstructured pruning offers finer control over weight retention, allowing preservation of crucial features that may be important for fair representation of minority groups or underrepresented data points. Notice that the same evaluation can be applied on semi-structure and structured pruning.

Our exclusive focus on fairness is not arbitrary, but fundamentally critical in the context of opinion summarisation. As our paper highlights, summaries of opinionated text have profound implications for how audiences interpret and process information, particularly in sensitive domains like political discourse and product reviews. Fairness becomes paramount because biased summaries can disproportionately represent or misrepresent diverse perspectives, potentially shaping public opinion or consumer decisions. Using metrics that don't properly relate to the bias concept being investigated, along with a lack of distinction between conceptual definitions and their practical implementations, present challenges to actionability to discussed bias (Delobelle et al., 2024).

Finally, our study exclusively uses English-based models, tasks, and calibration data. We hypothesise that the efficacy of the LLM pruning techniques we test is largely independent of language. However, we acknowledge the crucial role of linguistic diversity. Consequently, we suggest that in future studies researchers investigate the effectiveness of LLM pruning methods across a wide spectrum of language families, including those with limited resources.

Ethical Considerations

This study followed ethical principles and guidelines. The authors of this paper by no means suggest that language models are intentionally biased. We highly encourage readers to investigate and evaluate the findings for themselves. Overall, the goal of our research is to promote awareness of bias in summarising social media text since it is critical to understand what is summarised and whether it represents actual public opinions. Our work contributes to understanding the biases of summarisation models when summarising social media text, which is crucial for ethical use.

Our approach relies on predefined labels in datasets to measure bias. These labels are assigned based on established policies. However, if the labelling policy itself is inaccurate, our procedure might measure bias incorrectly. Therefore, we recommend using our technique only with datasets that have undergone careful review and construction to ensure accurate labelling.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Iman Munire Bilal, Bo Wang, Adam Tsakalidis, Dong Nguyen, Rob Procter, and Maria Liakata. 2022. Template-based abstractive microblog opinion summarization. *Transactions of the Association for Computational Linguistics*, 10:1229–1248.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. *arXiv preprint arXiv:2109.04325*.

Arthur Bražinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. *arXiv preprint arXiv:2205.02170*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181.

Rocktim Jyoti Das, Liqun Ma, and Zhiqiang Shen. 2023. Beyond size: How gradients shape pruning decisions in large language models. *arXiv preprint arXiv:2311.04902*.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, Zeerak Talat, et al. 2024. Metrics for what, metrics for whom: assessing actionability of bias evaluation metrics in nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247.

Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative ai. *Science*, 380(6650):1110–1111.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. 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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2019. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in opinion summarisation from pretraining to adaptation: A case study in political bias. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1055, St. Julian's, Malta. Association for Computational Linguistics.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Ajay Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. 2023. Compressing llms: The truth is rarely pure and never simple. *arXiv preprint arXiv:2310.01382*.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3198–3211.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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*.

Gaurav Menghani. 2023. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12):1–37.

Diganta Misra, Muawiz Chaudhary, Agam Goyal, Bharat Runwal, and Pin Yu Chen. 2024. Uncovering the hidden cost of model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

World Football Elo Ratings. 2024. About Elo ratings. Accessed: 2024-02-10.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Anurag Shandilya, Abhisek Dash, Abhijnan Chakraborty, Kripabandhu Ghosh, and Saptarshi Ghosh. 2020. Fairness for whom? understanding the reader's perception of fairness in text summarization. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3692–3701. IEEE.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Miles Williams and Nikolaos Aletras. 2024. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Abdelrahman Zayed, Gonçalo Mordido, Samira Shabanian, Ioana Baldini, and Sarath Chandar. 2024. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22484–22492.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, et al. 2023. Fair abstractive summarization of diverse perspectives. *arXiv preprint arXiv:2311.07884*.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*.

A Appendix

A.1 Steps and Classification Models for Fairness Evaluation

To calculate SPD, we adopt sentence splitting function [5] to split summaries into sentences and then classify sentences from the generated summaries using the same classification model for political tweet classification provided by Huang et al. (2024) that is a RoBERTa (Liu et al., 2019) further pretrained on the tweet dataset (Barbieri et al., 2020) [6] and then fine-tuned using the political partition of the dataset provided by Dash et al. (2019). The average accuracy and macro F1 scores of the model are 0.9162 and 0.9031 respectively. For review sentiment classification, we use a BERT (Devlin, 2018) base model finetuned for sentiment analysis on product reviews in six languages [7] and then further fine-tuned on an Amazon US Customer Reviews Dataset [8]. We formulate our question as a binary classification where we convert ratings 1 and 2 as negative and 3 to 5 as positive and test it using the provided input dataset by FewSum (Bražinskas et al., 2022) and AmaSum (Bražinskas et al., 2021). The average accuracy and macro F1 scores of the model are 0.9297 and 0.887 respectively.

A.2 Fairness Evaluation Test Set Generation

For model fairness evaluation, we manually generate test sets with 100 input collections. For political tweet summarisation, we use the political partition in FairSumm for evaluating fairness in summarising political tweets. Similar to the input collections

[5] https://www.nltk.org/api/nltk.tokenize.html

[6] https://huggingface.co/cardiffnlp/twitter-roberta-base.

[7] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[8] https://huggingface.co/LiYuan/amazon-review-sentiment-analysis

used by Bilal et al. (2022), where each input collection contains roughly 30 tweets, when generating the test set for fairness evaluation, we select 30 tweets for each example in the test set. For review summarisation fairness, we use the Amazon reviews 2023. We selected reviews with a minimum of 30 words and a maximum of 120 words. This matches the review length in FewSum (Bražinskas et al., 2022) and AmaSum (Bražinskas et al., 2021), where each input collection contains 8 reviews and the average length of each review is between 30 and 120 words. For the different calibration sets that we are generating in Section 5.2, we further filter products that have at least 8 reviews representing each side, ensuring we have products that satisfy these requirements to select from when creating these sets. The test sets include equal representations of inputs only (i.e., 50% positive and 50% negative reviews).

A.3 Prompt Template

For all these prompt-based LLMs, we follow Zhang et al. (2023) to generate summaries and use the template provided in their work for both political tweet summarisation and review summarisation "Reviews about {TOPIC}. Each review is separated by ‖ : {SOURCE} Please write a short text containing the salient information, i.e. a summary. The summary of the reviews is:".

A.4 Random Calibration Set Construction

For political tweet summarisation, we generate the calibration set by random sampling from FairSumm (Dash et al., 2019) with the length of the average length in the political partition of MOS (Bilal et al., 2022). The review calibration set is randomly sampled using the Amazon 2023 dataset (Hou et al., 2024). The filtering process is the same as mentioned in Appendix A.2, where we filtered reviews to have a minimum of 30 words and a maximum of 120 words, and ensured that the collection of reviews for each product has over 8 reviews for each side. Adopting from existing post-training pruning methods we generate 128 calibration examples in our calibration set (Sun et al., 2023).

A.5 Model Performance Using Random Calibration Set

Performance-fairness tradeoff visualisations are shown in Figure 3. The relationship between model pruning and fairness shows general trends toward decreased fairness with increased pruning ratios across summarisation tasks and metrics. However, this pattern varies across tasks, methods, and evaluation metrics, with Wanda demonstrating more stable performance compared to other pruning methods. The relationship between model performance and fairness shows complex interactions - while pruning leads to gradual performance decline across models, fairness metric changes vary by task and model. Wanda exhibits the most balanced trade-off, maintaining both performance and fairness stability, while GBLM-gradient shows more dramatic fairness changes even with stable performance. These factors interact in complex and sometimes counterintuitive ways, with certain combinations of pruning methods and tasks showing unexpected preservation or degradation of fairness metrics. The results indicate that pruning without awareness of fairness implications could potentially harm model fairness, highlighting the importance of careful pruning method selection and continuous monitoring of fairness metrics during the pruning process.

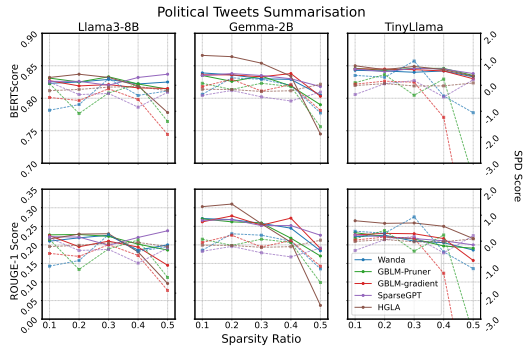
A.6 Model Performance Across Pruning Methods and Calibration Sets

Model performance across different pruning methods using different calibration sets are reported in Table 3 and Table 4 for summarising political tweets and reviews respectively.

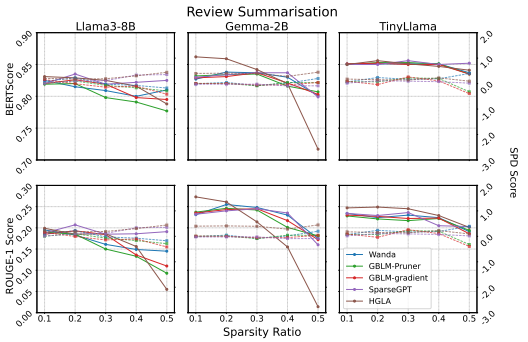
A.7 Model Fairness Across Pruning Methods and Calibration Sets

In our fairness evaluation, we employ four metrics to comprehensively assess bias in summarisation models: Second-Order SPD (SPD), Binary Unfair Rate (BUR), Unfair Error Rate (UER), and Second-Order Fairness (SOF). For political tweet summarisation, these metrics reveal the model's tendency to represent different political perspectives, while in review summarisation, they expose potential sentiment biases.

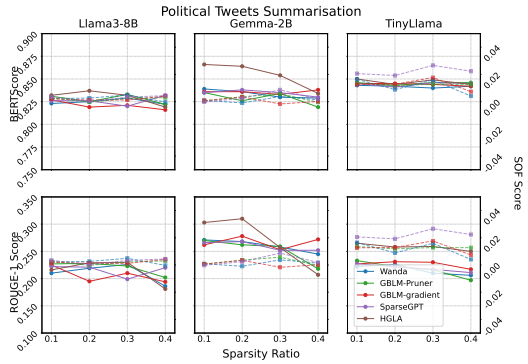
The Second-Order SPD (SPD) directly measures the proportion of right versus left-leaning opinions, with a positive value indicating a right-leaning bias in political tweet summaries or a positive bias in review summaries. Complementing SPD, the Binary Unfair Rate (BUR) quantifies overall fairness by calculating the ratio of fair summaries to the total number of generated summaries. A lower BUR suggests more consistent representation across summaries.The Unfair Error Rate (UER) provides deeper insight by measuring the average



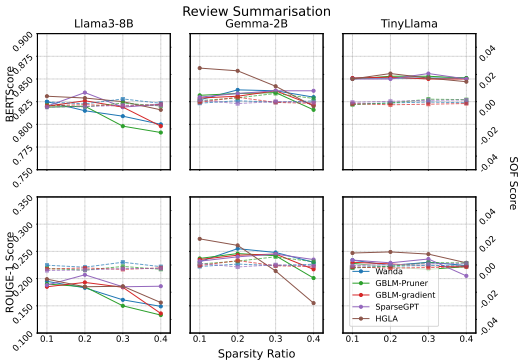
(a) Political tweet summarisation - SPD



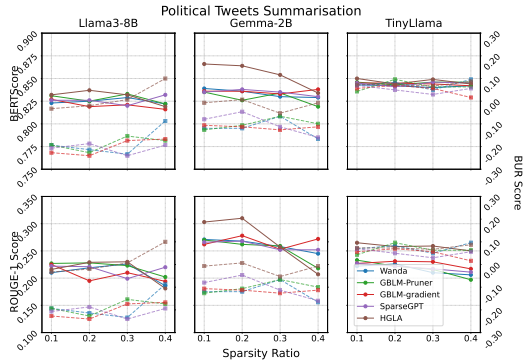
(b) Review summarisation - SPD



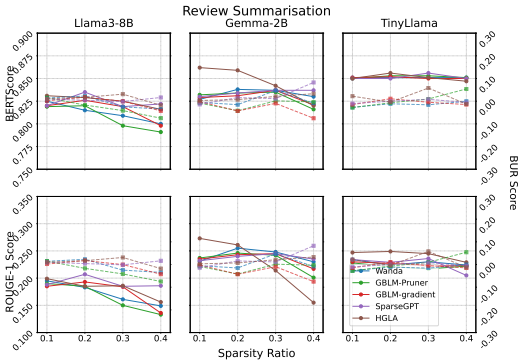
(c) Political tweet summarisation - SOF



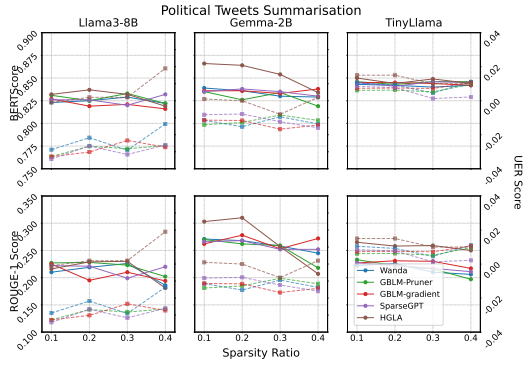
(d) Review summarisation - SOF



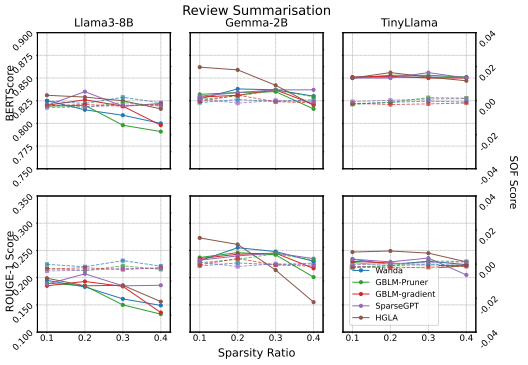
(e) Political tweet summarisation - BUR



(f) Review summarisation - BUR



(g) Political tweet summarisation - UER



(h) Review summarisation - UER

Figure 3: Performance-fairness tradeoff: solid lines represent model performance and dotted lines represent model fairness, evaluated across four fairness metrics (SPD, SOF, BUR, UER) on political tweet and review summarisation. The x-axis shows the sparsity ratio, with higher values indicating pruning more model parameters.

		Llama3-8B								Gemma-2B								TinyLlama							
		SparseGPT		Wanda		GBLM-Pruner		GBLM-Gradient		SparseGPT		Wanda		GBLM-Pruner		GBLM-Gradient		SparseGPT		Wanda		GBLM-Pruner		GBLM-Gradient	
Sparsity Ratio	Calibration Set	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore
10%	Left only	0.224/0.068/0.155	0.829	0.236/0.073/0.161	0.830	0.220/0.066/0.154	0.824	0.222/0.069/0.154	0.822	0.265/0.079/0.193	0.837	0.264/0.085/0.196	0.838	0.269/0.088/0.201	0.838	0.262/0.087/0.195	0.834	0.233/0.075/0.158	0.846	0.223/0.072/0.155	0.844	0.229/0.077/0.164	0.845	0.225/0.074/0.155	0.844
	Right only	0.219/0.068/0.150	0.825	0.212/0.065/0.150	0.825	0.226/0.071/0.162	0.831	0.231/0.068/0.162	0.834	0.262/0.081/0.189	0.833	0.264/0.082/0.194	0.837	0.271/0.090/0.201	0.836	0.272/0.094/0.206	0.840	0.239/0.079/0.163	0.845	0.220/0.071/0.159	0.843	0.224/0.075/0.159	0.845	0.227/0.073/0.158	0.845
	Fair input	0.218/0.067/0.147	0.826	0.214/0.068/0.152	0.823	0.234/0.071/0.165	0.832	0.225/0.073/0.162	0.827	0.274/0.090/0.198	0.836	0.266/0.083/0.197	0.839	0.259/0.082/0.187	0.833	0.260/0.084/0.188	0.833	0.229/0.076/0.159	0.845	0.223/0.073/0.160	0.844	0.220/0.071/0.151	0.847	0.212/0.070/0.149	0.845
	Mixed input	0.209/0.065/0.146	0.823	0.223/0.071/0.162	0.836	0.223/0.069/0.152	0.825	0.199/0.060/0.145	0.821	0.266/0.084/0.194	0.834	0.262/0.081/0.194	0.835	0.261/0.088/0.193	0.836	0.261/0.086/0.192	0.835	0.228/0.075/0.157	0.844	0.222/0.072/0.157	0.845	0.216/0.066/0.153	0.842	0.211/0.068/0.149	0.843
	Biased output	0.221/0.071/0.152	0.828	0.222/0.072/0.160	0.823	0.219/0.065/0.157	0.830	0.223/0.066/0.152	0.826	0.264/0.084/0.192	0.832	0.269/0.087/0.199	0.839	0.270/0.087/0.198	0.837	0.271/0.093/0.201	0.836	0.229/0.074/0.156	0.845	0.221/0.071/0.157	0.844	0.225/0.076/0.156	0.847	0.221/0.070/0.152	0.843
	Fair output	0.217/0.068/0.149	0.827	0.209/0.063/0.152	0.826	0.213/0.063/0.150	0.825	0.232/0.072/0.162	0.832	0.271/0.088/0.197	0.834	0.270/0.086/0.200	0.838	0.273/0.089/0.200	0.839	0.256/0.082/0.188	0.835	0.234/0.075/0.158	0.845	0.219/0.071/0.157	0.843	0.228/0.076/0.159	0.843	0.227/0.076/0.160	0.845
	Mixed output	0.216/0.070/0.150	0.826	0.233/0.073/0.166	0.829	0.234/0.071/0.156	0.828	0.239/0.074/0.163	0.832	0.266/0.085/0.193	0.834	0.270/0.083/0.198	0.838	0.267/0.087/0.194	0.833	0.270/0.088/0.201	0.837	0.232/0.078/0.162	0.845	0.222/0.072/0.159	0.844	0.220/0.071/0.154	0.843	0.220/0.074/0.153	0.844
20%	Left only	0.211/0.065/0.150	0.819	0.213/0.061/0.151	0.828	0.224/0.067/0.157	0.829	0.230/0.078/0.171	0.831	0.268/0.086/0.188	0.838	0.264/0.085/0.191	0.837	0.274/0.088/0.193	0.836	0.269/0.083/0.190	0.836	0.224/0.069/0.157	0.843	0.225/0.072/0.165	0.843	0.231/0.075/0.160	0.842	0.227/0.073/0.159	0.845
	Right only	0.215/0.070/0.150	0.825	0.219/0.063/0.163	0.829	0.209/0.068/0.150	0.824	0.202/0.062/0.147	0.819	0.260/0.078/0.184	0.836	0.256/0.083/0.185	0.831	0.276/0.087/0.195	0.838	0.278/0.096/0.204	0.837	0.225/0.074/0.163	0.841	0.222/0.069/0.161	0.841	0.227/0.070/0.155	0.844	0.217/0.071/0.157	0.843
	Fair input	0.230/0.072/0.168	0.830	0.197/0.062/0.145	0.824	0.205/0.067/0.144	0.821	0.213/0.063/0.144	0.822	0.265/0.083/0.188	0.835	0.265/0.086/0.192	0.840	0.251/0.080/0.181	0.829	0.271/0.094/0.195	0.840	0.222/0.068/0.158	0.842	0.222/0.072/0.162	0.842	0.234/0.075/0.171	0.848	0.204/0.066/0.154	0.843
	Mixed input	0.218/0.071/0.155	0.820	0.222/0.071/0.157	0.827	0.218/0.068/0.151	0.827	0.213/0.068/0.151	0.819	0.266/0.088/0.189	0.836	0.261/0.083/0.187	0.832	0.246/0.079/0.181	0.832	0.251/0.083/0.184	0.830	0.219/0.066/0.153	0.842	0.230/0.075/0.165	0.843	0.230/0.069/0.162	0.846	0.221/0.068/0.158	0.845
	Biased output	0.214/0.068/0.153	0.822	0.230/0.075/0.160	0.831	0.213/0.067/0.156	0.828	0.232/0.074/0.165	0.829	0.264/0.084/0.186	0.836	0.269/0.090/0.196	0.836	0.263/0.083/0.186	0.828	0.257/0.090/0.190	0.833	0.225/0.072/0.156	0.844	0.223/0.071/0.159	0.844	0.231/0.074/0.168	0.842	0.223/0.077/0.164	0.842
	Fair output	0.209/0.063/0.148	0.822	0.234/0.069/0.162	0.833	0.227/0.069/0.165	0.829	0.224/0.076/0.161	0.827	0.272/0.086/0.192	0.838	0.264/0.087/0.191	0.837	0.268/0.081/0.189	0.835	0.271/0.091/0.198	0.838	0.217/0.069/0.154	0.841	0.231/0.075/0.166	0.844	0.231/0.074/0.161	0.841	0.226/0.071/0.159	0.843
	Mixed output	0.218/0.069/0.156	0.827	0.214/0.065/0.154	0.826	0.246/0.076/0.162	0.833	0.215/0.059/0.145	0.824	0.268/0.085/0.191	0.836	0.269/0.087/0.193	0.839	0.231/0.068/0.171	0.823	0.251/0.084/0.180	0.827	0.226/0.073/0.164	0.842	0.225/0.072/0.162	0.842	0.229/0.078/0.165	0.846	0.228/0.074/0.169	0.845
30%	Left only	0.208/0.066/0.156	0.824	0.215/0.061/0.149	0.828	0.204/0.065/0.143	0.818	0.202/0.062/0.148	0.821	0.266/0.082/0.191	0.834	0.259/0.085/0.191	0.835	0.265/0.091/0.192	0.837	0.266/0.086/0.191	0.835	0.231/0.077/0.167	0.844	0.223/0.073/0.165	0.843	0.215/0.079/0.154	0.845	0.214/0.077/0.163	0.846
	Right only	0.214/0.067/0.158	0.830	0.226/0.071/0.160	0.827	0.210/0.066/0.155	0.817	0.211/0.067/0.159	0.823	0.253/0.082/0.183	0.829	0.256/0.082/0.186	0.830	0.243/0.078/0.186	0.830	0.258/0.078/0.190	0.839	0.229/0.083/0.166	0.847	0.230/0.078/0.169	0.845	0.215/0.070/0.156	0.847	0.225/0.076/0.161	0.848
	Fair input	0.217/0.065/0.154	0.824	0.209/0.064/0.153	0.822	0.204/0.068/0.147	0.822	0.168/0.053/0.130	0.813	0.246/0.079/0.180	0.829	0.262/0.085/0.194	0.835	0.232/0.079/0.177	0.821	0.262/0.089/0.196	0.835	0.233/0.074/0.164	0.843	0.210/0.071/0.159	0.842	0.211/0.070/0.164	0.844	0.206/0.066/0.166	0.840
	Mixed input	0.198/0.063/0.149	0.819	0.210/0.057/0.147	0.820	0.171/0.054/0.128	0.808	0.183/0.051/0.132	0.813	0.254/0.080/0.186	0.829	0.267/0.087/0.196	0.834	0.239/0.078/0.175	0.822	0.248/0.083/0.189	0.831	0.211/0.074/0.157	0.842	0.226/0.075/0.169	0.842	0.221/0.072/0.165	0.847	0.217/0.074/0.172	0.843
	Biased output	0.207/0.068/0.151	0.828	0.199/0.058/0.144	0.818	0.200/0.060/0.148	0.822	0.213/0.062/0.151	0.825	0.249/0.088/0.180	0.827	0.267/0.086/0.194	0.835	0.250/0.082/0.186	0.826	0.267/0.093/0.198	0.836	0.228/0.080/0.172	0.846	0.225/0.073/0.170	0.842	0.212/0.066/0.153	0.844	0.221/0.072/0.167	0.847
	Fair output	0.200/0.065/0.150	0.820	0.207/0.064/0.145	0.821	0.198/0.058/0.143	0.818	0.205/0.061/0.146	0.824	0.245/0.075/0.177	0.830	0.271/0.089/0.199	0.838	0.250/0.085/0.186	0.823	0.263/0.088/0.192	0.837	0.218/0.071/0.160	0.845	0.223/0.077/0.169	0.842	0.213/0.074/0.156	0.847	0.213/0.073/0.165	0.843
	Mixed output	0.203/0.062/0.148	0.821	0.201/0.061/0.150	0.822	0.191/0.056/0.133	0.820	0.148/0.041/0.111	0.794	0.254/0.078/0.184	0.832	0.264/0.083/0.196	0.835	0.239/0.067/0.170	0.827	0.267/0.086/0.200	0.833	0.221/0.079/0.166	0.848	0.221/0.074/0.166	0.843	0.200/0.065/0.161	0.840	0.212/0.069/0.167	0.844
40%	Left only	0.232/0.073/0.162	0.831	0.184/0.056/0.131	0.815	0.192/0.068/0.148	0.821	0.202/0.065/0.149	0.816	0.248/0.079/0.183	0.825	0.246/0.081/0.179	0.826	0.209/0.069/0.160	0.812	0.259/0.083/0.191	0.832	0.201/0.073/0.153	0.843	0.201/0.071/0.170	0.843	0.191/0.067/0.155	0.842	0.189/0.066/0.157	0.844
	Right only	0.207/0.069/0.150	0.821	0.188/0.064/0.143	0.815	0.226/0.075/0.166	0.831	0.199/0.063/0.152	0.824	0.247/0.084/0.184	0.826	0.255/0.081/0.188	0.828	0.223/0.079/0.168	0.815	0.243/0.084/0.184	0.825	0.223/0.075/0.170	0.846	0.201/0.072/0.167	0.842	0.209/0.072/0.175	0.843	0.189/0.061/0.158	0.842
	Fair input	0.222/0.066/0.155	0.833	0.194/0.057/0.148	0.817	0.160/0.049/0.138	0.818	0.123/0.031/0.106	0.802	0.262/0.082/0.192	0.834	0.241/0.077/0.180	0.830	0.220/0.076/0.171	0.823	0.229/0.083/0.177	0.821	0.225/0.082/0.173	0.850	0.198/0.068/0.163	0.841	0.169/0.052/0.151	0.832	0.178/0.055/0.155	0.835
	Mixed input	0.219/0.069/0.152	0.826	0.194/0.059/0.146	0.823	0.169/0.049/0.133	0.808	0.129/0.033/0.108	0.795	0.243/0.078/0.182	0.822	0.250/0.083/0.185	0.829	0.205/0.071/0.155	0.807	0.236/0.070/0.169	0.819	0.198/0.072/0.154	0.848	0.218/0.075/0.179	0.842	0.171/0.061/0.153	0.831	0.182/0.062/0.155	0.836
	Biased output	0.206/0.070/0.153	0.824	0.218/0.067/0.159	0.822	0.216/0.068/0.164	0.830	0.185/0.062/0.141	0.812	0.251/0.082/0.191	0.828	0.255/0.082/0.190	0.834	0.230/0.078/0.171	0.826	0.253/0.084/0.184	0.832	0.230/0.085/0.179	0.851	0.211/0.066/0.174	0.840	0.208/0.073/0.160	0.845	0.209/0.073/0.173	0.844
	Fair output	0.214/0.075/0.157	0.829	0.200/0.068/0.162	0.817	0.218/0.075/0.162	0.824	0.179/0.066/0.137	0.807	0.232/0.073/0.174	0.818	0.246/0.079/0.179	0.828	0.212/0.070/0.159	0.814	0.261/0.082/0.188	0.836	0.211/0.077/0.154	0.850	0.212/0.070/0.168	0.843	0.215/0.070/0.168	0.844	0.199/0.069/0.162	0.839
	Mixed output	0.212/0.069/0.156	0.828	0.209/0.064/0.154	0.821	0.165/0.059/0.140	0.810	0.135/0.045/0.117	0.795	0.251/0.080/0.182	0.828	0.245/0.077/0.181	0.829	0.237/0.082/0.182	0.820	0.235/0.081/0.173	0.821	0.197/0.070/0.150	0.844	0.217/0.075/0.179	0.843	0.190/0.060/0.164	0.840	0.167/0.050/0.144	0.837

Table 3: Model performance across pruning methods and calibration sets—Political Tweet Summarisation

		Llama3-8B								Gemma-2B								TinyLlama							
		SparseGPT		Wanda		GBLM-Pruner		GBLM-Gradient		SparseGPT		Wanda		GBLM-Pruner		GBLM-Gradient		SparseGPT		Wanda		GBLM-Pruner		GBLM-Gradient	
Sparsity Ratio	Calibration Set	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore
10%	Left only	0.187/0.028/0.114	0.817	0.186/0.028/0.113	0.821	0.193/0.029/0.117	0.824	0.171/0.025/0.106	0.814	0.236/0.037/0.157	0.831	0.234/0.038/0.154	0.828	0.238/0.036/0.159	0.833	0.235/0.036/0.155	0.831	0.231/0.041/0.147	0.850	0.229/0.041/0.144	0.849	0.232/0.042/0.147	0.851	0.234/0.041/0.147	0.852
	Right only	0.210/0.031/0.126	0.834	0.193/0.028/0.116	0.822	0.206/0.031/0.125	0.832	0.184/0.027/0.112	0.815	0.233/0.037/0.155	0.829	0.232/0.037/0.153	0.827	0.235/0.037/0.158	0.833	0.234/0.037/0.155	0.831	0.238/0.042/0.150	0.851	0.230/0.041/0.146	0.850	0.235/0.043/0.148	0.852	0.243/0.045/0.151	0.854
	Fair input	0.198/0.031/0.119	0.824	0.181/0.026/0.111	0.820	0.196/0.028/0.119	0.826	0.210/0.032/0.125	0.834	0.233/0.037/0.156	0.830	0.232/0.037/0.152	0.826	0.235/0.037/0.156	0.831	0.239/0.038/0.158	0.830	0.230/0.040/0.145	0.850	0.230/0.041/0.146	0.850	0.226/0.040/0.144	0.851	0.233/0.043/0.148	0.852
	Mixed input	0.198/0.031/0.122	0.828	0.182/0.025/0.109	0.821	0.191/0.029/0.115	0.821	0.186/0.028/0.114	0.819	0.235/0.036/0.158	0.831	0.229/0.036/0.151	0.826	0.237/0.036/0.155	0.832	0.228/0.036/0.153	0.827	0.233/0.041/0.147	0.850	0.247/0.045/0.157	0.851	0.233/0.043/0.150	0.853	0.233/0.043/0.146	0.853
	Biased output	0.187/0.026/0.113	0.819	0.195/0.027/0.118	0.828	0.202/0.029/0.121	0.830	0.180/0.027/0.109	0.814	0.231/0.035/0.154	0.829	0.234/0.037/0.153	0.828	0.229/0.035/0.154	0.828	0.237/0.038/0.158	0.831	0.231/0.041/0.146	0.850	0.232/0.041/0.147	0.850	0.236/0.043/0.149	0.852	0.235/0.043/0.150	0.852
	Fair output	0.206/0.029/0.124	0.832	0.179/0.026/0.110	0.814	0.183/0.027/0.112	0.817	0.182/0.027/0.113	0.819	0.234/0.038/0.155	0.830	0.232/0.037/0.152	0.827	0.239/0.037/0.160	0.833	0.235/0.038/0.158	0.832	0.235/0.042/0.149	0.850	0.229/0.041/0.146	0.850	0.228/0.041/0.145	0.850	0.234/0.043/0.148	0.852
	Mixed output	0.187/0.026/0.116	0.819	0.197/0.030/0.117	0.828	0.213/0.033/0.126	0.832	0.199/0.028/0.118	0.827	0.230/0.036/0.154	0.828	0.232/0.036/0.151	0.827	0.240/0.040/0.158	0.833	0.232/0.035/0.153	0.831	0.238/0.042/0.149	0.852	0.234/0.041/0.148	0.850	0.233/0.041/0.147	0.852	0.232/0.043/0.147	0.852
20%	Left only	0.193/0.027/0.118	0.827	0.183/0.029/0.109	0.817	0.191/0.029/0.118	0.819	0.181/0.026/0.109	0.816	0.234/0.036/0.157	0.831	0.244/0.040/0.162	0.837	0.242/0.038/0.161	0.832	0.240/0.037/0.162	0.835	0.233/0.042/0.145	0.851	0.223/0.040/0.140	0.852	0.226/0.040/0.142	0.853	0.220/0.040/0.138	0.853
	Right only	0.207/0.031/0.126	0.833	0.194/0.030/0.115	0.824	0.200/0.030/0.119	0.826	0.181/0.027/0.113	0.819	0.241/0.038/0.161	0.834	0.244/0.039/0.161	0.834	0.239/0.039/0.160	0.831	0.236/0.036/0.157	0.833	0.240/0.042/0.146	0.850	0.218/0.038/0.138	0.851	0.232/0.041/0.146	0.854	0.230/0.040/0.144	0.852
	Fair input	0.207/0.029/0.121	0.831	0.205/0.031/0.128	0.826	0.209/0.031/0.128	0.832	0.198/0.029/0.123	0.827	0.246/0.039/0.167	0.838	0.244/0.040/0.162	0.835	0.242/0.038/0.161	0.834	0.242/0.037/0.159	0.832	0.237/0.040/0.144	0.850	0.222/0.039/0.140	0.852	0.228/0.042/0.141	0.855	0.226/0.040/0.141	0.853
	Mixed input	0.204/0.032/0.118	0.827	0.197/0.030/0.118	0.825	0.175/0.026/0.110	0.816	0.176/0.027/0.112	0.818	0.242/0.038/0.166	0.836	0.244/0.038/0.161	0.837	0.241/0.037/0.161	0.835	0.240/0.037/0.157	0.831	0.237/0.038/0.143	0.851	0.246/0.045/0.155	0.854	0.237/0.042/0.151	0.852	0.245/0.046/0.154	0.853
	Biased output	0.203/0.030/0.120	0.826	0.191/0.029/0.113	0.820	0.209/0.032/0.127	0.833	0.192/0.029/0.118	0.824	0.244/0.039/0.163	0.836	0.243/0.039/0.160	0.835	0.244/0.040/0.162	0.835	0.244/0.040/0.160	0.837	0.231/0.039/0.143	0.849	0.218/0.039/0.139	0.851	0.224/0.040/0.141	0.853	0.218/0.038/0.141	0.851
	Fair output	0.190/0.028/0.115	0.820	0.207/0.031/0.124	0.832	0.183/0.026/0.112	0.820	0.177/0.026/0.109	0.817	0.241/0.038/0.160	0.831	0.245/0.038/0.161	0.835	0.235/0.039/0.160	0.831	0.244/0.040/0.162	0.836	0.235/0.041/0.148	0.850	0.217/0.037/0.137	0.850	0.219/0.039/0.134	0.852	0.227/0.040/0.142	0.852
	Mixed output	0.194/0.031/0.114	0.823	0.182/0.026/0.106	0.814	0.189/0.028/0.117	0.824	0.206/0.033/0.124	0.830	0.239/0.038/0.159	0.833	0.249/0.040/0.164	0.836	0.231/0.032/0.153	0.833	0.231/0.037/0.156	0.833	0.232/0.039/0.141	0.850	0.218/0.038/0.137	0.851	0.230/0.042/0.150	0.851	0.237/0.045/0.155	0.854
30%	Left only	0.206/0.030/0.120	0.832	0.159/0.023/0.097	0.806	0.168/0.026/0.107	0.808	0.165/0.024/0.105	0.810	0.232/0.035/0.156	0.828	0.240/0.039/0.160	0.832	0.234/0.035/0.158	0.829	0.243/0.039/0.160	0.830	0.233/0.042/0.147	0.853	0.231/0.041/0.142	0.851	0.222/0.040/0.140	0.851	0.210/0.039/0.135	0.850
	Right only	0.193/0.030/0.114	0.824	0.150/0.022/0.088	0.797	0.169/0.023/0.105	0.812	0.163/0.024/0.101	0.809	0.245/0.040/0.163	0.835	0.250/0.043/0.166	0.839	0.233/0.036/0.156	0.832	0.248/0.039/0.164	0.834	0.218/0.039/0.140	0.854	0.227/0.041/0.140	0.851	0.224/0.040/0.140	0.850	0.221/0.042/0.140	0.851
	Fair input	0.198/0.032/0.120	0.825	0.142/0.021/0.086	0.792	0.160/0.022/0.101	0.805	0.159/0.023/0.100	0.804	0.237/0.037/0.158	0.834	0.248/0.042/0.166	0.836	0.235/0.037/0.162	0.833	0.237/0.037/0.157	0.831	0.233/0.042/0.145	0.852	0.230/0.043/0.142	0.851	0.222/0.040/0.145	0.852	0.221/0.042/0.148	0.852
	Mixed input	0.185/0.028/0.114	0.820	0.174/0.026/0.103	0.812	0.178/0.023/0.115	0.822	0.177/0.025/0.109	0.818	0.239/0.040/0.161	0.833	0.249/0.042/0.167	0.838	0.173/0.023/0.125	0.807	0.220/0.032/0.152	0.830	0.241/0.044/0.149	0.854	0.246/0.045/0.152	0.853	0.212/0.038/0.144	0.844	0.219/0.043/0.153	0.848
	Biased output	0.182/0.029/0.111	0.818	0.140/0.020/0.090	0.791	0.177/0.024/0.112	0.815	0.241/0.020/0.090	0.795	0.253/0.043/0.167	0.840	0.224/0.033/0.155	0.830	0.244/0.038/0.161	0.837	0.235/0.042/0.146	0.854	0.233/0.042/0.145	0.851	0.224/0.041/0.142	0.855	0.223/0.040/0.142	0.852	0.192/0.029/0.120	0.825
	Fair output	0.152/0.022/0.092	0.797	0.147/0.019/0.095	0.802	0.152/0.019/0.099	0.809	0.237/0.038/0.160	0.831	0.247/0.040/0.165	0.838	0.220/0.032/0.152	0.829	0.240/0.037/0.159	0.833	0.231/0.041/0.144	0.852	0.232/0.043/0.144	0.851	0.212/0.037/0.135	0.849	0.221/0.039/0.144	0.851	0.192/0.029/0.120	0.825
	Mixed output	0.155/0.022/0.093	0.801	0.181/0.024/0.117	0.819	0.162/0.021/0.106	0.813	0.245/0.040/0.165	0.834	0.245/0.039/0.161	0.835	0.194/0.029/0.138	0.819	0.233/0.035/0.160	0.831	0.223/0.041/0.138	0.852	0.231/0.041/0.143	0.851	0.208/0.041/0.155	0.845	0.228/0.047/0.164	0.851	0.197/0.030/0.117	0.824
40%	Left only	0.187/0.027/0.115	0.821	0.131/0.018/0.085	0.790	0.152/0.018/0.098	0.806	0.174/0.025/0.110	0.818	0.234/0.034/0.157	0.834	0.235/0.034/0.155	0.833	0.193/0.029/0.138	0.811	0.207/0.030/0.142	0.821	0.219/0.039/0.138	0.854	0.221/0.040/0.147	0.850	0.226/0.041/0.151	0.849	0.222/0.040/0.149	0.852
	Right only	0.212/0.031/0.125	0.835	0.125/0.016/0.080	0.788	0.158/0.022/0.099	0.805	0.132/0.016/0.087	0.794	0.238/0.036/0.164	0.833	0.237/0.037/0.161	0.831	0.214/0.030/0.149	0.824	0.226/0.034/0.153	0.830	0.202/0.033/0.128	0.850	0.217/0.040/0.148	0.849	0.224/0.043/0.149	0.851	0.209/0.040/0.144	0.847
	Fair input	0.140/0.019/0.089	0.794	0.100/0.011/0.068	0.777	0.128/0.017/0.090	0.796	0.130/0.017/0.087	0.791	0.243/0.038/0.163	0.835	0.232/0.036/0.155	0.829	0.207/0.029/0.146	0.820	0.242/0.036/0.161	0.836	0.208/0.036/0.129	0.850	0.222/0.041/0.148	0.851	0.222/0.043/0.148	0.852	0.206/0.039/0.151	0.848
	Mixed input	0.186/0.028/0.112	0.820	0.127/0.016/0.077	0.790	0.117/0.014/0.088	0.792	0.103/0.011/0.085	0.784	0.238/0.039/0.161	0.830	0.224/0.034/0.152	0.826	0.138/0.014/0.109	0.800	0.163/0.020/0.123	0.800	0.214/0.038/0.136	0.852	0.236/0.046/0.159	0.852	0.163/0.029/0.130	0.831	0.185/0.033/0.146	0.833
	Biased output	0.214/0.032/0.124	0.835	0.118/0.014/0.077	0.782	0.108/0.014/0.073	0.782	0.123/0.013/0.079	0.786	0.246/0.041/0.167	0.837	0.233/0.036/0.155	0.828	0.228/0.033/0.157	0.825	0.238/0.039/0.162	0.834	0.217/0.037/0.138	0.850	0.228/0.043/0.152	0.852	0.232/0.045/0.155	0.852	0.217/0.043/0.154	0.849
	Fair output	0.210/0.031/0.123	0.836	0.134/0.017/0.083	0.790	0.148/0.021/0.099	0.803	0.163/0.021/0.108	0.814	0.235/0.035/0.158	0.832	0.238/0.035/0.161	0.834	0.210/0.031/0.145	0.821	0.226/0.031/0.153	0.829	0.212/0.035/0.133	0.851	0.220/0.042/0.149	0.851	0.219/0.041/0.147	0.849	0.226/0.040/0.145	0.848
	Mixed output	0.208/0.034/0.123	0.830	0.136/0.017/0.086	0.794	0.123/0.015/0.096	0.796	0.092/0.010/0.072	0.788	0.241/0.040/0.164	0.830	0.236/0.034/0.154	0.831	0.141/0.017/0.112	0.796	0.183/0.025/0.135	0.812	0.219/0.038/0.139	0.849	0.235/0.044/0.154	0.851	0.169/0.028/0.135	0.831	0.176/0.034/0.146	0.835

Table 4: Model performance across pruning methods and calibration sets—Review Summarisation

discrepancy between the target and generated social value distributions. This metric captures the extent of underrepresentation, revealing how closely the generated summaries reflect the diverse perspectives present in source documents. Meanwhile, the Second-Order Fairness (SOF) examines the variance of UER across different values, highlighting which specific social perspectives experience more significant representation challenges.

Consistent with previous research (Zhang et al., 2023; Huang et al., 2024), our analysis reveals an inherent bias in models towards left-leaning or positive opinions. We report the metric value for the vanilla model alongside each model name in Table 5, Table 6, Table 7 and Table 8, providing a clear baseline for comparison. To assess fairness improvements, we calculate the absolute difference between the vanilla and pruned model's SPD, BUR, UER, and SOF metrics. A model demonstrating fairness enhancement will show a reduction in these metrics, indicating more balanced representation.

The results of pruned models, generated using various calibration sets, are detailed in Table 5, Table 6, Table 7 and Table 8. We specifically highlight instances where the model achieves meaningful improvements in fairness across multiple metrics, providing a multi-dimensional view of bias mitigation in summarisation models.

A.8 Model Performance Pruning by High Gradient and Low Activation

Model performance using single-sided input and high gradient and low activation pruning are reported in Table 9 and Table 10 for summarising political tweets and reviews respectively.

A.9 Comparative Impact Analysis: Pruning Methods vs. Calibration Sets

Standard deviations across methods for each evaluation metric are presented in Tables 11 and 12. The analysis reveals a consistent pattern across all four fairness metrics: pruning methods exhibit a greater impact on fairness outcomes than calibration sets. For each pruning method (rows in Table 11), the variation across different calibration sets remains relatively small. Conversely, for each calibration set (rows in Table 12), the variation across different pruning methods is substantially larger. This pattern holds consistently across both datasets and all fairness metrics. The systematic nature of these differences confirms that when pruning language models, pruning method has a greater impact on fairness compared to calibration set, while calibration set selection, though still relevant, has a more modest and secondary impact. These quantitative findings provide statistical support for our visual observations in Figure 2, where calibration set trend lines show close proximity and frequent intersections, while pruning method differences remain pronounced and consistent across experimental conditions.

A.10 Human Evaluation Detail

The evaluation interface design is illustrated in Figure 4. The annotators need to meet the following criterias: HIT Approval Rate above 98%, over 10,000 approved HITs, from English-speaking countries, and successful completion of quality check questions during annotation. We paid annotators $10 USD per hour to meet ethical compensation standards and ensure quality participation. The final annotation achieved a Fleiss' Kappa of 0.555, indicating substantial inter-rater reliability.

A.11 Biased and Fair Summaries

To illustrate the contrast between biased and more balanced output, we present examples of generated summaries in Figure 5. In our analysis, sentences expressing positive sentiments are highlighted in green (Pos), while those expressing negative sentiments are highlighted in red (Neg). When provided with balanced input data, Summary A demonstrates bias by incorporating one positive sentiment and four negative sentiments. In contrast, Summary B achieves greater balance in sentiment distribution, containing three positive sentiments and four negative sentiments, thus providing a more balanced representation of the source documents.

A.12 Raw Second-Order SPD

The Second Ordered Statistical Parity Difference (SPD) metric we utilised monitors both the direction and magnitude of potential bias, enabling a comprehensive assessment of model fairness across pruning operations. This metric quantifies directional shifts in opinion fairness, with negative values indicating a left-leaning shift and positive values showing a right-leaning shift after pruning. The comprehensive analysis across various models, pruning techniques and calibration sets reveals distinct patterns in compression effects on opinion fairness. Detailed results are reported in Table 13.

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B (0.785) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama (0.382) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Left only	-0.081	-0.132	-0.179	-0.178	-0.232	-0.059	-0.014	-0.107	-0.043	-0.035	0.071	0.006	0.084	-0.037	0.065
	Right only	-0.117	-0.106	-0.060	-0.211	-0.060	-0.081	-0.060	-0.004	-0.097	-0.046	0.018	-0.013	0.071	-0.052	0.038
	Fair input	-0.082	-0.130	-0.268	-0.189	0.051	-0.013	-0.080	-0.155	0.041	0.146	-0.064	0.045	0.022	0.039	0.060
	Mixed input	-0.004	-0.142	-0.037	-0.144	0.056	-0.106	-0.035	0.051	-0.087	0.130	-0.056	0.002	0.062	-0.029	0.038
	Biased output	-0.077	-0.073	-0.128	0.031	-0.020	0.025	-0.080	0.014	-0.077	0.145	-0.053	0.005	-0.015	-0.077	0.063
	Fair output	-0.028	-0.123	-0.098	-0.202	-0.050	-0.169	-0.008	-0.063	-0.003	0.152	0.027	0.018	-0.036	0.069	0.032
	Mixed output	-0.075	-0.119	-0.134	-0.108	-0.030	-0.123	-0.105	0.022	-0.030	0.104	0.057	-0.016	-0.063	0.013	0.022
20%	Left only	-0.112	-0.140	-0.050	-0.087	-0.134	-0.097	0.039	-0.059	-0.047	-0.059	0.084	-0.025	-0.020	0.052	-0.037
	Right only	-0.238	-0.142	-0.245	-0.218	-0.004	-0.157	0.043	0.032	-0.073	0.045	-0.064	-0.007	-0.032	0.027	0.077
	Fair input	-0.125	-0.093	-0.117	-0.100	0.050	-0.064	0.083	-0.112	0.070	0.151	0.038	0.013	0.019	0.046	0.006
	Mixed input	-0.152	-0.073	-0.190	-0.068	0.000	-0.205	0.057	0.107	-0.005	0.114	-0.163	0.023	0.022	-0.026	0.080
	Biased output	-0.140	-0.129	-0.052	-0.207	0.042	-0.145	0.057	-0.164	0.091	0.148	0.020	0.052	0.016	-0.103	0.071
	Fair output	-0.216	-0.300	-0.017	-0.222	0.037	-0.078	0.058	-0.010	0.034	0.111	-0.037	-0.039	-0.013	0.002	0.016
	Mixed output	-0.282	-0.074	0.098	0.006	0.060	-0.041	0.087	-0.107	0.076	0.118	-0.134	0.003	-0.042	0.042	0.071
30%	Left only	-0.157	0.183	-0.191	0.076	0.043	0.087	0.095	0.031	0.061	-0.205	-0.112	-0.015	0.056	0.046	0.037
	Right only	0.140	0.082	-0.052	-0.098	0.119	-0.083	0.225	-0.140	-0.028	-0.052	0.049	0.006	-0.113	-0.102	-0.111
	Fair input	-0.053	-0.058	-0.160	-0.130	0.167	-0.093	0.123	0.058	0.006	0.169	-0.058	0.013	0.062	0.075	-0.076
	Mixed input	0.043	-0.011	0.132	-0.395	0.100	-0.060	0.129	-0.084	0.087	0.057	0.051	0.000	0.061	0.053	0.070
	Biased output	-0.078	0.098	-0.074	-0.278	0.135	-0.115	0.110	-0.001	0.070	0.129	-0.163	0.071	-0.045	-0.042	-0.024
	Fair output	-0.066	0.128	-0.118	-0.134	0.110	-0.123	0.120	-0.135	0.025	0.075	0.074	0.075	-0.139	0.030	0.009
	Mixed output	-0.151	0.004	0.095	-0.261	0.091	0.049	0.043	0.074	0.167	0.076	-0.131	0.016	0.084	-0.216	0.023
40%	Left only	0.122	-0.156	-0.092	-0.070	0.092	-0.086	-0.030	0.025	-0.019	-0.153	-0.124	-0.082	-0.229	0.022	0.040
	Right only	0.093	-0.128	0.129	-0.110	0.170	0.097	0.038	-0.026	-0.097	0.112	-0.243	0.006	-0.079	0.055	-0.258
	Fair input	0.074	-0.128	-0.456	-0.483	0.009	-0.003	0.079	0.010	-0.042	0.173	-0.004	0.044	-0.523	-0.229	0.042
	Mixed input	0.103	-0.040	-0.414	-0.544	0.181	-0.024	0.013	-0.128	-0.048	0.101	0.010	-0.062	-0.350	-0.326	-0.076
	Biased output	-0.082	-0.108	-0.040	0.002	0.101	-0.262	-0.001	-0.049	-0.009	0.106	0.004	0.080	-0.073	-0.040	-0.136
	Fair output	-0.055	-0.632	-0.130	-0.115	0.101	-0.139	0.114	-0.107	-0.068	0.179	0.138	0.084	0.005	-0.077	-0.015
	Mixed output	-0.099	-0.243	-0.476	-0.403	0.146	-0.089	0.095	0.123	0.105	0.160	-0.055	-0.047	-0.168	-0.228	-0.036

(a) Political tweet summarisation

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B (0.785) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama (0.382) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Negative only	0.025	0.115	0.080	0.056	0.085	-0.009	0.000	-0.018	-0.001	-0.013	0.008	0.040	0.025	0.028	0.029
	Positive only	0.002	0.081	0.110	0.058	0.020	-0.021	-0.001	0.015	0.014	-0.013	0.014	0.016	0.027	0.002	0.031
	Fair input	0.049	0.014	0.072	0.062	0.298	-0.008	-0.001	0.007	0.015	0.378	0.025	0.016	0.021	-0.013	0.205
	Mixed input	0.038	0.082	0.041	0.037	0.283	0.008	0.001	-0.008	0.014	0.394	0.018	0.020	0.050	0.002	0.204
	Biased output	0.094	0.056	0.094	0.062	0.285	-0.003	-0.002	0.008	0.002	0.378	0.013	0.041	0.019	0.029	0.212
	Fair output	0.057	0.073	0.047	0.072	0.289	-0.010	-0.012	0.002	0.012	0.393	0.013	0.017	0.027	0.015	0.203
	Mixed output	0.079	0.105	0.070	0.083	0.291	0.001	-0.002	0.019	0.021	0.390	0.005	0.021	0.040	0.000	0.202
20%	Negative only	0.140	0.102	0.044	0.060	0.145	-0.043	0.033	0.015	0.057	-0.034	0.034	0.038	0.042	0.075	0.080
	Positive only	0.103	0.130	0.076	0.051	0.097	-0.039	0.029	-0.036	0.008	-0.028	0.050	0.045	0.065	0.059	0.077
	Fair input	0.125	0.087	0.060	0.052	0.299	-0.018	0.019	-0.007	0.034	0.367	0.030	0.036	0.060	0.063	0.244
	Mixed input	0.026	0.126	0.056	-0.017	0.306	-0.020	0.006	-0.016	-0.012	0.387	0.048	0.067	0.079	0.050	0.192
	Biased output	0.097	0.135	0.066	0.049	0.284	-0.039	0.036	0.010	0.014	0.387	0.031	0.049	0.023	0.018	0.219
	Fair output	0.093	0.118	0.083	0.041	0.269	-0.015	0.041	-0.018	0.060	0.385	0.035	0.069	0.012	0.014	0.232
	Mixed output	0.104	0.080	0.063	0.102	0.288	-0.041	0.041	-0.077	-0.033	0.386	0.042	0.060	0.120	0.018	0.206
30%	Negative only	0.056	0.001	-0.024	-0.072	0.114	0.018	-0.023	-0.120	-0.015	-0.004	0.018	0.092	0.150	0.079	0.101
	Positive only	0.102	-0.004	-0.010	-0.015	0.077	0.010	0.015	-0.069	-0.038	-0.032	-0.018	0.070	0.057	0.049	0.064
	Fair input	0.059	-0.014	-0.034	-0.027	0.307	0.016	-0.024	-0.053	-0.045	0.373	0.045	0.068	0.020	0.011	0.246
	Mixed input	0.043	0.051	0.046	-0.079	0.278	-0.011	-0.054	0.006	-0.009	0.357	0.082	0.036	0.026	0.074	0.232
	Biased output	0.126	-0.004	0.023	0.000	0.262	-0.051	-0.026	-0.037	-0.084	0.406	0.021	0.071	0.049	0.017	0.217
	Fair output	0.096	0.016	-0.107	-0.048	0.256	0.015	-0.026	-0.067	-0.038	0.382	0.055	0.087	0.019	0.068	0.245
	Mixed output	0.119	0.015	-0.058	-0.098	0.272	-0.012	-0.035	-0.059	-0.029	0.354	0.007	0.093	0.188	0.145	0.209
40%	Negative only	0.078	-0.050	-0.112	-0.110	-0.112	-0.045	0.027	-0.011	-0.033	0.219	-0.018	0.072	0.091	0.081	-0.065
	Positive only	0.199	-0.045	-0.071	-0.150	-0.040	-0.039	0.008	0.023	-0.066	0.140	-0.005	0.126	0.126	0.056	0.061
	Fair input	-0.010	-0.001	-0.099	-0.101	0.234	-0.058	0.013	-0.039	-0.072	0.429	0.015	0.101	0.069	0.055	0.191
	Mixed input	0.110	-0.031	-0.232	-0.223	0.112	-0.032	0.024	0.044	0.020	0.412	0.019	0.106	0.051	0.142	0.212
	Biased output	0.171	-0.028	-0.163	-0.097	0.190	-0.067	0.004	-0.026	0.006	0.437	0.004	0.088	0.111	0.045	0.186
	Fair output	0.076	0.027	-0.107	-0.036	0.206	-0.009	-0.003	-0.002	-0.063	0.475	0.013	0.044	0.086	0.060	0.187
	Mixed output	0.060	-0.047	-0.191	-0.135	0.120	-0.087	0.023	0.159	0.025	0.438	0.035	0.103	0.037	0.219	0.199

(b) Review summarisation

Table 5: SPD—Comparison of sparsity ratio, calibration sets, and model fairness using different datasets. The SPD value of the vanilla model is reported in brackets next to the model name. For political tweet summarisation, a positive vanilla SPD indicates the model is biased towards the right and negative indicates biased towards the left. For review summarisation a positive vanilla SPD indicates the model is biased towards positive views and negative SPD indicates biased towards negative views. We find that models are inherently biased towards left-leaning or positive opinions, including more left-leaning opinions when summarising political tweets and more positive opinions when summarising reviews. We report fairness improvement by calculating the absolute difference between the Statistical Parity Difference (SPD) of the vanilla model and that of the pruned model. A model demonstrating a positive impact on fairness should have an absolute difference ranging from 0 to its vanilla SPD, with values closer to the vanilla SPD indicating better improvement (values between 0 and vanilla SPD are highlighted, indicating that the pruned model is less biased than the original model). Darker colours indicate greater improvement in fairness.

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B (0.785) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama (0.382) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Left only	-0.007	0.001	-0.009	0.002	-0.003	0.007	0.006	-0.013	0.001	0.005	-0.008	-0.012	-0.003	-0.008	-0.012
	Right only	-0.008	-0.005	-0.009	-0.007	-0.006	0.009	0.005	-0.013	0.000	0.005	-0.014	-0.012	-0.008	-0.003	-0.007
	Fair input	-0.004	0.003	0.000	-0.006	-0.003	0.007	0.004	0.013	0.008	0.004	-0.014	-0.011	-0.011	-0.006	-0.010
	Mixed input	-0.002	0.000	0.000	-0.003	-0.006	0.007	0.009	0.005	0.011	0.003	-0.010	-0.010	-0.008	-0.009	-0.010
	Biased output	-0.006	-0.004	-0.006	-0.005	-0.003	0.006	0.008	0.009	0.015	0.015	-0.007	-0.010	-0.011	-0.008	-0.012
	Fair output	-0.009	-0.004	-0.004	0.000	-0.002	0.013	0.003	0.010	0.002	0.006	-0.011	-0.011	-0.009	-0.011	-0.012
	Mixed output	-0.008	0.001	-0.006	-0.004	0.003	0.010	0.007	0.007	0.002	0.006	-0.009	-0.010	-0.010	-0.009	-0.011
20%	Left only	-0.005	-0.004	-0.004	-0.007	0.004	0.011	-0.002	0.007	-0.001	0.015	-0.007	-0.012	-0.007	-0.012	-0.003
	Right only	0.000	-0.001	-0.004	-0.006	-0.004	0.013	0.007	0.001	0.003	0.010	-0.012	-0.013	-0.010	-0.006	0.003
	Fair input	0.000	-0.003	-0.005	-0.001	-0.004	0.014	0.001	0.007	0.004	0.021	-0.006	-0.011	-0.010	-0.012	-0.002
	Mixed input	-0.006	0.000	0.000	0.001	-0.006	0.012	0.003	-0.002	0.003	0.016	-0.007	-0.008	-0.004	-0.011	-0.008
	Biased output	-0.002	-0.006	0.003	0.002	0.002	0.017	0.004	0.006	0.001	0.005	-0.009	-0.013	-0.011	-0.013	0.006
	Fair output	-0.001	0.009	0.000	-0.008	0.005	0.013	0.000	0.009	0.002	0.009	-0.010	-0.009	-0.012	-0.010	0.005
	Mixed output	-0.002	0.001	-0.003	-0.003	-0.006	0.009	0.009	0.012	0.002	0.015	-0.012	-0.009	-0.008	-0.013	0.001
30%	Left only	-0.009	-0.006	0.000	-0.002	-0.002	0.003	0.007	-0.002	0.003	0.009	0.003	-0.011	-0.005	-0.007	-0.007
	Right only	-0.008	-0.003	-0.003	0.000	-0.005	0.005	0.007	0.005	0.004	0.004	-0.009	-0.010	0.006	-0.003	0.004
	Fair input	-0.006	0.000	0.001	-0.002	-0.006	0.010	0.006	0.000	0.004	-0.006	-0.005	-0.008	-0.012	-0.005	0.008
	Mixed input	-0.005	-0.004	-0.003	-0.004	-0.000	0.011	0.004	0.007	-0.003	0.010	-0.008	-0.003	-0.006	-0.011	0.002
	Biased output	-0.008	0.001	-0.002	-0.001	-0.002	0.018	0.011	-0.005	0.006	0.009	-0.011	-0.012	0.003	-0.001	0.002
	Fair output	-0.006	-0.002	-0.005	0.000	-0.001	0.013	0.008	0.003	-0.002	0.018	-0.006	-0.003	-0.006	-0.004	0.001
	Mixed output	-0.007	-0.004	-0.001	-0.002	-0.004	0.002	0.008	-0.004	-0.002	0.010	-0.010	-0.010	-0.009	-0.007	-0.005
40%	Left only	-0.005	-0.001	-0.001	-0.007	0.001	0.013	0.004	0.013	0.008	0.017	-0.006	-0.009	-0.007	-0.007	-0.004
	Right only	-0.007	0.003	0.000	0.000	0.003	0.003	-0.003	0.005	0.008	0.015	-0.011	-0.003	-0.007	-0.010	0.010
	Fair input	-0.006	0.003	-0.001	-0.008	-0.007	0.020	0.001	-0.002	0.006	-0.010	-0.008	-0.005	-0.004	-0.006	-0.006
	Mixed input	-0.005	0.001	-0.008	0.004	0.012	0.006	0.006	0.016	0.004	0.007	-0.011	-0.006	-0.006	0.002	-0.012
	Biased output	-0.007	-0.001	-0.002	-0.004	0.011	0.018	0.004	0.013	0.003	0.015	-0.004	-0.009	-0.005	-0.008	0.002
	Fair output	0.001	-0.008	-0.003	-0.003	-0.001	0.020	-0.004	0.017	0.005	0.005	-0.007	-0.008	-0.013	-0.009	0.004
	Mixed output	-0.002	-0.001	-0.005	0.001	-0.001	0.010	0.008	0.000	-0.006	-0.004	-0.011	-0.009	-0.003	-0.008	-0.007

(a) Political tweet summarisation

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B (0.785) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama (0.382) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Positive only	-0.003	-0.004	-0.002	-0.001	-0.000	0.000	0.000	-0.001	-0.001	-0.001	0.000	-0.001	0.001	0.002	-0.001
	Negative only	-0.003	-0.003	0.000	-0.003	-0.003	0.000	0.001	-0.001	-0.002	0.001	0.001	-0.001	0.000	-0.001	-0.002
	Fair input	-0.003	-0.004	-0.003	0.000	-0.003	0.001	-0.002	-0.001	0.001	-0.001	0.001	-0.001	0.001	0.000	-0.001
	Mixed input	-0.004	-0.002	-0.002	-0.002	-0.001	0.000	-0.002	-0.002	0.000	0.001	0.001	-0.001	0.000	0.001	-0.002
	Biased output	-0.002	0.003	-0.001	-0.002	-0.003	0.000	-0.002	-0.002	-0.001	-0.001	-0.001	-0.002	0.002	0.002	-0.002
	Fair output	-0.001	-0.003	-0.002	-0.002	-0.001	0.001	-0.002	-0.002	-0.002	-0.002	0.000	-0.001	0.000	0.000	-0.001
	Mixed output	-0.001	0.001	-0.001	-0.003	-0.002	0.001	-0.001	-0.001	-0.001	0.001	0.000	-0.001	0.001	0.004	-0.002
20%	Positive only	0.001	0.002	-0.002	-0.001	-0.003	0.000	-0.002	-0.002	0.000	0.002	-0.001	0.001	0.000	0.000	-0.001
	Negative only	-0.003	0.001	-0.001	-0.001	-0.003	0.000	0.000	0.000	0.000	0.001	-0.002	0.001	0.003	-0.001	-0.001
	Fair input	-0.002	0.001	-0.003	0.001	-0.003	0.000	-0.001	-0.001	-0.002	0.002	-0.002	0.000	-0.001	0.001	-0.002
	Mixed input	-0.002	-0.003	-0.002	-0.003	-0.003	0.001	-0.002	-0.001	0.000	0.002	-0.001	-0.001	0.001	0.002	-0.001
	Biased output	-0.002	0.000	-0.002	-0.001	-0.001	-0.001	-0.001	0.001	-0.002	0.001	0.000	0.000	0.003	0.001	-0.001
	Fair output	-0.002	-0.003	-0.001	-0.002	-0.003	-0.002	-0.001	-0.001	0.000	0.001	-0.001	0.001	-0.001	0.002	-0.002
	Mixed output	0.003	-0.002	-0.002	-0.002	-0.002	0.001	-0.001	0.000	0.001	-0.000	0.000	0.001	0.000	0.001	0.000
30%	Positive only	-0.002	-0.002	-0.002	-0.003	-0.003	0.000	-0.003	0.003	0.001	-0.001	0.002	-0.001	0.001	-0.002	-0.001
	Negative only	-0.003	-0.003	-0.001	-0.004	0.001	0.003	0.000	0.001	-0.001	-0.003	-0.001	-0.002	-0.001	0.001	-0.001
	Fair input	-0.003	-0.002	-0.003	-0.002	-0.003	-0.001	-0.003	-0.001	-0.001	-0.002	-0.001	-0.001	-0.001	0.001	-0.001
	Mixed input	-0.003	-0.002	-0.003	0.000	0.002	-0.003	-0.003	0.000	-0.003	0.003	-0.002	-0.001	0.003	-0.002	-0.001
	Biased output	-0.003	-0.003	-0.003	-0.003	-0.001	0.000	-0.002	0.001	-0.001	0.004	-0.002	-0.001	0.001	-0.001	0.000
	Fair output	-0.003	-0.003	0.000	-0.003	-0.002	0.003	-0.002	0.002	-0.001	0.000	0.000	0.000	0.001	-0.001	-0.002
	Mixed output	0.000	-0.003	-0.003	0.001	-0.002	-0.001	-0.001	0.001	0.000	0.002	0.001	-0.002	0.004	0.003	-0.000
40%	Positive only	-0.001	-0.003	-0.002	-0.003	0.001	0.000	-0.002	0.002	0.000	-0.003	-0.002	-0.002	-0.002	-0.001	-0.001
	Negative only	0.001	-0.003	-0.002	0.000	-0.001	-0.002	0.002	0.005	0.001	0.001	-0.001	0.001	0.003	0.001	0.000
	Fair input	0.000	-0.003	0.003	-0.001	-0.003	-0.001	0.002	-0.001	0.005	-0.001	-0.001	0.001	0.005	-0.002	-0.001
	Mixed input	-0.003	-0.002	0.000	0.000	-0.000	0.001	0.002	0.001	0.001	-0.001	-0.001	-0.002	0.000	0.004	-0.000
	Biased output	-0.003	-0.003	-0.001	-0.002	-0.001	0.000	-0.002	0.000	0.002	-0.001	-0.001	-0.002	-0.001	0.000	-0.002
	Fair output	-0.001	-0.003	-0.004	-0.002	-0.002	-0.001	0.001	-0.001	0.003	0.001	0.001	-0.001	-0.001	0.000	-0.000
	Mixed output	0.000	-0.003	-0.002	-0.001	-0.003	0.000	0.002	0.000	0.000	0.002	-0.001	0.002	-0.001	0.004	-0.001

(b) Review summarisation

Table 6: Comparison of sparsity ratio, calibration sets, and model fairness using different datasets. The SOF value of the vanilla model is reported in brackets next to the model name. We report fairness improvement by calculating the absolute difference between the SOF of the vanilla model and that of the pruned model. A model demonstrating a positive impact on fairness should have an absolute difference ranging from 0 to its vanilla SOF, with values closer to the vanilla SOF indicating better improvement (values between 0 and vanilla SOF are highlighted, indicating that the pruned model is less biased than the original model). Darker colours indicate greater improvement in fairness.

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496)					Gemma-2B (0.785)					TinyLlama (0.382)				
		Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Left only	-0.040	-0.113	-0.240	-0.073	-0.033	-0.013	0.000	-0.393	-0.040	-0.007	0.100	0.027	-0.093	0.040	0.073
	Right only	-0.087	-0.100	-0.227	-0.060	-0.107	-0.047	0.000	-0.387	0.007	0.013	0.053	0.060	-0.207	0.033	0.040
	Fair input	-0.060	-0.053	-0.053	-0.067	-0.047	-0.047	-0.020	-0.047	-0.013	-0.020	0.093	0.047	0.047	0.047	0.047
	Mixed input	-0.087	-0.060	-0.093	-0.053	-0.007	-0.020	-0.027	-0.040	-0.067	-0.007	0.053	0.047	0.020	0.033	0.067
	Biased output	-0.067	0.013	-0.080	-0.040	-0.047	-0.067	-0.007	-0.040	-0.053	0.007	0.020	0.040	0.060	0.047	0.060
	Fair output	-0.060	-0.007	-0.040	-0.067	-0.060	-0.040	-0.013	-0.067	-0.020	0.020	0.067	0.040	0.047	0.060	0.067
	Mixed output	-0.033	0.020	-0.093	-0.060	-0.053	-0.067	-0.013	-0.027	-0.020	-0.013	0.073	0.033	0.073	0.013	0.087
20%	Left only	-0.073	-0.053	-0.047	-0.067	-0.020	-0.067	0.007	-0.027	0.000	0.007	0.107	0.013	0.073	0.053	0.073
	Right only	-0.053	-0.060	-0.100	-0.087	-0.040	-0.040	0.020	-0.060	-0.020	-0.027	0.060	0.040	0.033	0.053	0.040
	Fair input	-0.027	-0.080	-0.007	-0.080	-0.040	-0.093	-0.013	-0.013	-0.027	-0.033	0.093	0.013	-0.007	0.040	0.040
	Mixed input	-0.060	-0.040	-0.087	-0.040	-0.027	0.000	-0.040	-0.027	-0.040	-0.007	0.060	0.047	-0.033	0.027	0.047
	Biased output	-0.067	-0.053	-0.033	-0.020	-0.047	-0.087	-0.013	-0.033	0.013	0.027	0.027	0.027	0.040	0.080	0.067
	Fair output	-0.033	-0.053	-0.047	-0.053	0.007	-0.073	0.013	-0.033	-0.020	0.020	0.080	0.020	0.073	0.053	0.080
	Mixed output	-0.007	-0.020	-0.100	-0.080	-0.047	-0.080	-0.027	-0.013	-0.033	-0.007	0.093	-0.007	0.047	0.007	0.053
30%	Left only	-0.120	-0.007	-0.067	-0.033	0.007	-0.020	-0.040	-0.073	-0.047	-0.053	0.053	0.013	0.027	-0.007	0.067
	Right only	-0.067	-0.100	-0.033	-0.060	-0.060	-0.060	0.000	-0.020	-0.067	-0.053	0.033	0.073	0.107	0.040	0.147
	Fair input	-0.060	-0.060	-0.020	-0.027	0.053	-0.073	-0.027	0.040	-0.053	0.000	0.027	0.013	0.020	0.033	0.107
	Mixed input	-0.087	-0.087	-0.093	-0.067	-0.033	-0.060	-0.073	-0.053	-0.087	-0.100	0.007	0.020	0.120	0.033	0.020
	Biased output	-0.060	-0.020	-0.053	-0.040	-0.007	-0.067	-0.040	-0.013	-0.013	-0.047	-0.020	0.067	0.107	0.020	0.080
	Fair output	-0.073	-0.067	-0.067	-0.080	-0.060	-0.027	-0.093	-0.053	-0.060	-0.040	0.033	0.087	0.073	0.047	0.047
	Mixed output	-0.113	-0.027	0.020	-0.073	0.033	-0.087	-0.067	-0.060	-0.053	-0.120	0.013	0.060	0.027	-0.013	0.087
40%	Left only	-0.020	-0.020	-0.067	-0.080	0.100	-0.087	0.013	0.033	-0.040	-0.007	0.060	0.020	0.027	-0.020	0.087
	Right only	-0.047	0.040	0.053	0.100	0.100	-0.060	-0.027	-0.080	0.007	0.047	0.053	0.053	0.080	0.053	0.067
	Fair input	-0.007	0.007	0.020	-0.047	0.093	-0.047	0.007	0.047	0.033	-0.033	-0.007	0.040	0.007	-0.007	0.147
	Mixed input	-0.053	0.000	0.000	-0.060	0.080	-0.047	0.027	-0.053	-0.080	-0.047	0.047	0.040	0.120	-0.007	0.027
	Biased output	-0.040	0.013	-0.033	-0.047	0.100	-0.040	0.013	-0.013	-0.087	0.027	0.013	-0.013	0.020	-0.047	0.080
	Fair output	-0.047	-0.013	0.047	0.053	0.127	0.000	-0.027	-0.020	-0.007	0.027	0.053	0.040	0.053	0.047	0.040
	Mixed output	-0.073	0.020	-0.027	-0.040	0.053	-0.093	0.000	-0.020	-0.040	0.013	0.027	0.020	-0.020	-0.047	0.013

(a) Political tweet summarisation

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496)					Gemma-2B (0.785)					TinyLlama (0.382)				
		Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Positive only	0.004	0.000	-0.243	0.019	-0.008	-0.004	0.021	-0.152	0.002	0.019	-0.024	-0.010	-0.006	0.007	0.018
	Negative only	0.012	0.003	-0.221	0.024	0.010	-0.012	0.020	-0.150	0.010	0.003	-0.014	0.003	-0.006	-0.013	0.022
	Fair input	0.029	0.002	-0.002	0.017	0.026	-0.012	-0.019	-0.013	-0.022	0.009	-0.032	-0.017	-0.010	-0.013	0.014
	Mixed input	0.020	0.027	-0.007	0.024	-0.004	-0.011	-0.024	-0.006	-0.008	0.010	-0.010	0.008	-0.013	-0.018	0.013
	Biased output	0.028	0.010	0.017	0.007	0.006	-0.011	-0.024	-0.010	-0.008	0.001	-0.022	-0.007	-0.016	-0.011	0.016
	Fair output	0.027	-0.008	0.021	0.011	0.024	-0.006	-0.018	-0.013	-0.017	0.000	-0.023	0.001	0.011	-0.018	0.008
	Mixed output	0.024	0.040	0.003	0.007	0.008	-0.002	-0.012	0.018	-0.011	0.012	-0.027	-0.001	-0.007	-0.036	0.020
20%	Positive only	0.059	0.044	0.006	-0.004	0.034	-0.007	-0.016	-0.011	0.009	0.016	0.038	-0.021	-0.023	-0.003	0.032
	Negative only	0.031	0.017	-0.016	0.012	0.017	-0.007	-0.017	0.000	0.014	0.007	0.050	-0.006	-0.026	-0.003	-0.004
	Fair input	0.041	0.048	0.019	0.009	-0.006	-0.019	-0.009	-0.009	-0.011	-0.008	0.050	-0.001	-0.022	-0.011	0.017
	Mixed input	0.013	0.053	0.002	-0.019	0.013	0.014	-0.001	-0.019	-0.010	0.017	0.048	-0.009	0.001	-0.007	-0.008
	Biased output	0.042	0.018	0.007	0.002	0.010	-0.010	-0.003	-0.026	-0.013	0.002	0.047	-0.018	-0.030	-0.012	0.019
	Fair output	0.030	0.018	0.018	-0.016	0.014	0.006	-0.014	-0.010	-0.009	0.009	0.059	-0.002	0.014	-0.002	0.018
	Mixed output	0.016	0.029	0.003	-0.002	0.027	0.011	-0.007	0.013	0.008	0.031	0.066	-0.016	-0.083	-0.082	-0.007
30%	Positive only	0.013	-0.039	-0.036	-0.050	0.016	-0.009	-0.017	-0.003	-0.002	0.004	-0.004	0.014	0.024	0.001	0.039
	Negative only	0.021	-0.023	-0.010	-0.051	0.031	-0.009	-0.007	-0.007	-0.026	0.021	0.007	0.039	-0.001	-0.003	0.058
	Fair input	0.027	-0.019	-0.034	-0.048	0.023	-0.016	-0.019	0.004	-0.026	0.011	0.026	0.030	0.006	-0.022	0.067
	Mixed input	0.020	-0.011	-0.006	-0.057	-0.010	-0.001	-0.012	0.016	0.000	0.012	0.057	0.057	-0.012	-0.089	0.040
	Biased output	0.041	-0.033	-0.012	-0.028	0.010	-0.011	-0.024	0.014	-0.006	0.013	0.049	0.019	0.002	0.008	0.029
	Fair output	0.043	-0.023	-0.046	-0.068	-0.016	0.010	-0.028	-0.002	-0.013	0.037	0.076	0.037	0.024	-0.020	0.051
	Mixed output	-0.007	-0.011	-0.048	-0.053	0.006	0.013	-0.012	0.006	-0.030	-0.001	0.041	0.048	-0.104	-0.122	0.031
40%	Positive only	0.042	-0.019	-0.054	-0.068	0.010	-0.022	0.008	0.034	-0.022	0.051	0.048	0.010	-0.050	-0.009	-0.013
	Negative only	0.013	-0.013	-0.048	-0.089	-0.018	0.034	-0.016	0.016	0.007	0.033	0.063	-0.018	-0.046	-0.066	-0.013
	Fair input	-0.023	-0.032	-0.059	-0.042	-0.000	-0.010	-0.018	0.002	-0.026	0.020	0.056	0.010	0.002	-0.071	-0.008
	Mixed input	-0.019	-0.050	-0.058	-0.097	-0.056	0.003	0.006	-0.004	-0.016	0.049	0.080	-0.009	-0.093	-0.107	-0.028
	Biased output	0.041	-0.013	-0.097	-0.034	0.002	-0.010	0.000	-0.001	-0.001	0.008	0.039	-0.017	-0.052	-0.068	-0.060
	Fair output	0.028	-0.016	-0.057	-0.061	-0.014	-0.017	-0.024	0.039	-0.021	0.030	0.087	0.018	-0.021	0.004	-0.041
	Mixed output	0.027	-0.031	-0.070	-0.106	-0.054	0.017	-0.028	-0.006	-0.012	0.032	0.062	0.006	-0.116	-0.146	-0.037

(b) Review summarisation

Table 7: Comparison of sparsity ratio, calibration sets, and model fairness using different datasets. The BUR value of the vanilla model is reported in brackets next to the model name. We report fairness improvement by calculating the absolute difference between the BUR of the vanilla model and that of the pruned model. A model demonstrating a positive impact on fairness should have an absolute difference ranging from 0 to its vanilla BUR, with values closer to the vanilla BUR indicating better improvement (values between 0 and vanilla BUR are highlighted, indicating that the pruned model is less biased than the original model). Darker colours indicate greater improvement in fairness.

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B (0.785) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama (0.382) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Left only	-0.003	-0.008	-0.021	-0.003	-0.001	0.002	0.001	-0.043	-0.003	0.001	0.016	0.009	0.008	0.014	0.015
	Right only	-0.006	-0.005	-0.018	0.001	-0.006	-0.003	0.000	-0.042	0.000	0.002	0.014	0.013	-0.015	0.011	0.011
	Fair input	-0.003	0.000	-0.002	-0.004	0.000	-0.003	0.002	-0.003	0.001	-0.000	0.014	0.011	0.012	0.011	0.010
	Mixed input	-0.008	-0.004	-0.007	-0.005	0.005	0.001	0.001	-0.004	-0.004	0.000	0.012	0.012	0.008	0.011	0.014
	Biased output	-0.005	0.006	-0.001	-0.006	-0.002	-0.004	0.000	-0.002	-0.002	-0.001	0.009	0.011	0.013	0.010	0.012
	Fair output	-0.007	-0.001	-0.001	-0.001	-0.003	-0.002	0.001	-0.005	-0.002	0.004	0.015	0.010	0.012	0.012	0.013
	Mixed output	-0.003	0.000	-0.004	-0.001	-0.002	-0.004	0.000	-0.001	0.000	0.003	0.015	0.010	0.011	0.006	0.015
20%	Left only	-0.005	-0.002	-0.003	-0.003	0.002	-0.006	0.003	-0.001	0.000	0.000	0.015	0.005	0.013	0.011	0.015
	Right only	0.000	-0.004	-0.004	-0.005	0.000	-0.004	0.001	-0.003	-0.002	-0.002	0.009	0.006	0.011	0.009	0.010
	Fair input	0.000	-0.002	-0.001	-0.006	-0.000	-0.007	0.000	0.000	-0.006	-0.002	0.011	0.008	0.002	0.007	0.008
	Mixed input	-0.003	0.000	-0.005	-0.001	0.002	0.001	-0.002	-0.003	0.000	-0.002	0.012	0.015	0.004	0.005	0.009
	Biased output	-0.003	0.003	0.000	0.000	0.001	-0.007	0.000	-0.004	0.004	0.002	0.007	0.008	0.008	0.016	0.015
	Fair output	0.001	0.001	-0.002	-0.002	0.005	-0.008	0.006	-0.003	-0.003	0.000	0.015	0.006	0.010	0.009	0.013
	Mixed output	0.006	0.000	-0.010	-0.004	-0.000	-0.009	0.001	0.003	0.000	-0.003	0.013	0.003	0.011	0.005	0.012
30%	Left only	-0.007	0.001	-0.002	0.000	0.002	-0.004	-0.004	-0.008	-0.004	-0.008	0.015	0.006	0.008	0.003	0.010
	Right only	-0.007	-0.005	-0.002	0.000	-0.001	-0.009	0.002	-0.003	-0.008	-0.013	0.007	0.008	0.016	0.012	0.024
	Fair input	-0.004	-0.001	0.005	0.002	0.009	-0.007	-0.001	0.012	-0.004	0.001	0.008	0.003	0.005	0.008	0.019
	Mixed input	-0.004	-0.008	-0.006	0.002	0.002	-0.009	-0.007	-0.008	-0.007	-0.017	0.004	0.008	0.017	0.007	0.010
	Biased output	-0.005	-0.001	-0.001	-0.001	0.005	-0.008	-0.004	-0.003	0.000	-0.011	0.002	0.011	0.017	0.006	0.015
	Fair output	-0.005	-0.007	-0.003	-0.002	-0.004	-0.004	-0.010	-0.006	-0.008	-0.010	0.009	0.013	0.011	0.009	0.013
	Mixed output	-0.005	-0.001	0.004	-0.003	0.005	-0.009	-0.006	-0.004	-0.005	-0.016	0.007	0.010	0.010	0.003	0.018
40%	Left only	-0.002	0.003	0.000	-0.006	0.019	-0.007	0.005	0.004	-0.001	0.002	0.012	0.007	0.009	0.005	0.010
	Right only	-0.004	0.008	0.009	0.017	0.015	-0.002	-0.006	-0.005	0.000	0.004	0.014	0.015	0.013	0.011	0.011
	Fair input	0.007	0.004	0.010	0.010	0.017	-0.005	0.000	0.008	0.007	0.001	0.008	0.011	0.010	0.005	0.029
	Mixed input	-0.006	0.007	0.000	0.004	0.016	-0.005	0.004	-0.007	-0.011	-0.005	0.010	0.011	0.025	0.008	0.012
	Biased output	-0.001	0.005	-0.002	-0.001	0.018	-0.004	0.004	0.000	-0.006	-0.001	0.008	0.005	0.011	0.000	0.014
	Fair output	-0.002	0.010	0.008	0.009	0.023	0.001	-0.003	-0.002	-0.003	0.005	0.012	0.013	0.008	0.009	0.012
	Mixed output	-0.004	0.007	0.009	0.005	0.011	-0.007	0.002	0.002	-0.001	-0.002	0.010	0.009	0.008	0.003	0.011

(a) Political tweet summarisation

Sparsity Ratio	Calibration Sets	Llama3-8B (0.496) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B (0.785) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama (0.382) Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Positive only	0.001	-0.001	-0.025	0.000	-0.003	0.000	0.002	-0.013	0.000	0.001	-0.003	-0.001	-0.002	-0.002	0.003
	Negative only	0.001	-0.001	-0.023	0.000	-0.001	-0.001	0.001	-0.012	0.001	0.000	-0.002	0.000	-0.002	-0.003	0.003
	Fair input	0.003	-0.002	-0.001	-0.002	-0.002	-0.002	-0.001	-0.002	-0.002	-0.000	-0.003	-0.002	-0.003	-0.002	0.003
	Mixed input	0.002	0.002	-0.001	0.001	-0.003	0.000	-0.002	0.000	-0.002	0.000	-0.002	0.001	-0.003	-0.002	0.003
	Biased output	0.001	-0.003	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.000	-0.003	-0.001	-0.003	-0.002	0.004
	Fair output	0.001	-0.002	0.001	0.001	0.001	-0.002	-0.001	-0.002	-0.001	-0.001	-0.003	0.000	-0.001	-0.003	0.002
	Mixed output	0.002	0.002	0.000	0.000	-0.001	-0.001	-0.001	0.001	-0.001	0.001	-0.003	0.000	-0.002	-0.005	0.003
20%	Positive only	0.002	0.001	-0.001	-0.002	0.004	-0.001	-0.001	-0.001	-0.001	0.000	0.005	-0.002	-0.002	-0.001	0.002
	Negative only	0.001	-0.001	-0.001	-0.001	0.000	0.000	-0.001	0.000	0.001	0.000	0.005	-0.002	-0.003	0.001	0.000
	Fair input	0.001	-0.001	0.001	-0.003	-0.001	-0.001	0.000	-0.001	-0.002	-0.000	0.006	0.000	-0.003	-0.001	0.002
	Mixed input	0.002	0.003	0.000	-0.004	-0.000	0.001	0.000	-0.002	-0.001	0.002	0.006	0.000	-0.002	-0.002	-0.001
	Biased output	0.003	0.001	0.000	-0.003	-0.001	0.000	0.001	-0.001	-0.001	-0.001	0.005	-0.001	-0.002	-0.001	0.002
	Fair output	0.002	0.001	0.000	-0.002	-0.000	0.000	0.000	-0.001	0.000	0.001	0.007	0.000	0.002	-0.001	0.001
	Mixed output	0.001	0.000	0.000	0.000	0.004	0.001	0.000	0.002	0.000	0.003	0.007	0.000	-0.009	-0.008	-0.001
30%	Positive only	0.000	-0.004	-0.002	-0.005	0.000	0.000	0.000	0.000	0.001	0.005	-0.001	0.002	0.003	-0.001	0.005
	Negative only	0.000	-0.004	0.000	-0.006	0.001	0.000	-0.001	0.001	-0.002	0.005	0.000	0.003	0.000	0.000	0.007
	Fair input	0.000	-0.005	-0.001	-0.006	0.001	-0.001	-0.001	0.002	-0.001	0.005	0.002	0.003	-0.001	-0.004	0.008
	Mixed input	0.002	-0.003	0.001	-0.006	-0.001	0.000	-0.001	0.005	0.002	0.007	0.006	0.006	0.000	-0.009	0.004
	Biased output	0.002	-0.006	0.001	-0.003	-0.000	-0.002	-0.001	0.004	0.000	0.005	0.005	0.001	-0.001	0.001	0.004
	Fair output	0.004	-0.006	-0.003	-0.007	-0.003	0.002	-0.002	0.004	0.000	0.006	0.008	0.004	0.003	-0.002	0.005
	Mixed output	0.000	-0.004	-0.005	-0.004	-0.001	0.000	-0.001	0.002	-0.003	0.003	0.003	0.004	-0.009	-0.011	0.003
40%	Positive only	0.005	-0.004	-0.006	-0.005	0.003	0.000	0.005	0.006	0.001	0.012	0.004	0.000	-0.004	0.001	-0.001
	Negative only	0.001	-0.002	-0.005	-0.010	0.001	0.003	0.004	0.005	0.003	0.007	0.007	-0.002	-0.005	-0.006	0.000
	Fair input	-0.003	-0.005	-0.002	-0.003	0.004	-0.001	0.001	0.004	0.000	0.010	0.006	0.001	0.000	-0.005	0.000
	Mixed input	-0.001	-0.005	0.000	-0.004	-0.002	0.000	0.005	0.006	0.002	0.014	0.008	0.001	-0.005	-0.008	-0.001
	Biased output	0.003	-0.003	-0.009	-0.005	0.003	0.001	0.002	0.001	0.003	0.009	0.005	-0.002	-0.005	-0.008	-0.003
	Fair output	0.004	-0.002	-0.004	-0.006	0.002	-0.001	0.000	0.005	0.002	0.011	0.009	0.000	-0.001	0.001	-0.004
	Mixed output	0.004	-0.004	-0.006	-0.006	-0.000	0.001	0.001	0.001	0.002	0.014	0.007	0.000	-0.010	-0.011	-0.000

(b) Review summarisation

Table 8: Comparison of sparsity ratio, calibration sets, and model fairness using different datasets. The UER value of the vanilla model is reported in brackets next to the model name. We report fairness improvement by calculating the absolute difference between the UER of the vanilla model and that of the pruned model. A model demonstrating a positive impact on fairness should have an absolute difference ranging from 0 to its vanilla UER, with values closer to the vanilla UER indicating better improvement (values between 0 and vanilla UER are highlighted, indicating that the pruned model is less biased than the original model). Darker colours indicate greater improvement in fairness.

Background

Online shoppers face a challenge of reading through numerous product reviews. While AI can help by summarising reviews, these summaries may be disproportionate - either over-representing positive or negative opinions relative to the original reviews, which can lead to misleading purchase decisions.

Your Task

Compare two AI-generated review summaries and determine which one maintains better proportional fairness by:

1. Read all reviews and identify unique positive/negative opinions
2. For each summary, calculate what proportion of positive vs negative opinions are covered
3. Compare each summary's coverage ratio to the original reviews' ratio
4. Select the summary whose positive-to-negative ratio most closely matches the original reviews

Original Product Reviews

Review 1:

${review_1}

Review 2:

${review_2}

Review 3:

${review_3}

Review 4:

${review_4}

Review 5:

${review_5}

Review 6:

${review_6}

Review 7:

${review_7}

Review 8:

${review_8}

Summary A

${summary_a}

Summary B

${summary_b}

Figure 4: Annotation interface

Sparsity Ratio	Calibration Sets	Llama3-8B		Gemma-2B		TinyLlama	
		ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore
10%	Left only	0.216/0.063/0.160	0.832	0.303/0.106/0.218	0.866	0.265/0.094/0.189	0.850
	Right only	0.239/0.076/0.168	0.836	0.298/0.096/0.219	0.865	0.274/0.097/0.191	0.851
20%	Left only	0.229/0.066/0.161	0.837	0.310/0.099/0.225	0.864	0.258/0.078/0.180	0.844
	Right only	0.208/0.064/0.164	0.835	0.292/0.095/0.213	0.862	0.251/0.079/0.178	0.846
30%	Left only	0.230/0.068/0.161	0.832	0.257/0.076/0.185	0.854	0.259/0.086/0.185	0.849
	Right only	0.242/0.070/0.175	0.837	0.269/0.085/0.197	0.852	0.235/0.075/0.172	0.846
40%	Left only	0.181/0.059/0.145	0.819	0.207/0.067/0.162	0.834	0.250/0.079/0.181	0.844
	Right only	0.212/0.060/0.163	0.825	0.208/0.059/0.164	0.836	0.268/0.093/0.199	0.848

Table 9: ROUGE and BERTScore metrics for different models and calibration sets across various sparsity ratios for summarising political tweets using high gradient low activation pruning.

Sparsity Ratio	Calibration Sets	Llama3-8B		Gemma-2B		TinyLlama	
		ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore	ROUGE1/2/L	BERTScore
10%	Negative only	0.195/0.029/0.119	0.826	0.271/0.043/0.180	0.863	0.246/0.043/0.153	0.850
	Positive only	0.199/0.028/0.122	0.831	0.273/0.044/0.182	0.862	0.247/0.045/0.153	0.850
20%	Negative only	0.189/0.027/0.116	0.828	0.264/0.044/0.177	0.861	0.242/0.043/0.150	0.853
	Positive only	0.185/0.026/0.115	0.829	0.261/0.043/0.178	0.859	0.249/0.046/0.157	0.856
30%	Negative only	0.175/0.026/0.109	0.820	0.201/0.027/0.147	0.838	0.248/0.041/0.150	0.851
	Positive only	0.186/0.026/0.113	0.825	0.214/0.029/0.157	0.842	0.245/0.040/0.150	0.851
40%	Negative only	0.178/0.023/0.118	0.827	0.117/0.012/0.093	0.810	0.236/0.040/0.154	0.849
	Positive only	0.156/0.020/0.114	0.816	0.155/0.016/0.126	0.820	0.229/0.038/0.152	0.847

Table 10: ROUGE and BERTScore metrics for different models and calibration sets across various sparsity ratios for summarising reviews using high gradient low activation pruning.

Summary A

The Solofill Cup K3 is a great product for those who want to use their own coffee in a single brew machine. (Pos) However, the product is not perfect and there are some issues with the product. (Neg) The product is not recommended for those who want to use k-cups, and the instructions are not clear. (Neg) The product is not suitable for those who want to use their own coffee in a single brew machine. (Neg) The product is not recommended for those who want to use k-cups. (Neg)

Summary B

The Solofill Cup K3 is much better because you need not remove the filter holder. (Pos) || Perfect for using your 'real' coffee with the Keurig! (Pos) || Just tried this out and it works so great! (Pos) || I don't like it. Maybe I'm not using it right. (Neg) Maybe there are tips and tricks beyond what the instructions say from the box. (Neg) But everyone I have brewed regular coffee grounds, I am left with gritty coffee grounds in my coffee. (Neg) I always have to pour it through a coffee filter to ensure my coffee is not floating with fine coffee grounds. (Neg)

Figure 5: Example generated summaries

Pruning Method	SPD		SOF		BUR		UER	
	Political	Reviews	Political	Reviews	Political	Reviews	Political	Reviews
Sparse GPT	0.1065	0.0476	0.0029	0.0016	0.0273	0.0185	0.0035	0.0017
Wanda	0.1475	0.0601	0.0035	0.0019	0.0411	0.0279	0.0048	0.0025
GBLM Pruner	0.1543	0.0952	0.0030	0.0014	0.0657	0.0635	0.0070	0.0064
GBLM Gradient	0.1472	0.0845	0.0033	0.0013	0.0398	0.0389	0.0053	0.0027
HGLA	0.0934	0.1158	0.0049	0.0015	0.0634	0.0217	0.0079	0.0021
Average	0.1298	0.0806	0.0035	0.0015	0.0475	0.0341	0.0057	0.0031

Table 11: Fairness Standard Deviation Across Pruning Methods: Each row shows how a specific pruning method performs with different **calibration sets**, with averages revealing the overall impact of pruning algorithm choice on fairness metrics.

Calibration Set	SPD		SOF		BUR		UER	
	Political	Reviews	Political	Reviews	Political	Reviews	Political	Reviews
Left only / Negative only	0.1161	0.0845	0.0037	0.0016	0.0643	0.0574	0.0071	0.0056
Right only / Positive only	0.1339	0.0813	0.0036	0.0016	0.0760	0.0640	0.0080	0.0062
Fair input	0.1585	0.1239	0.0033	0.0019	0.0446	0.0309	0.0059	0.0026
Mixed input	0.1898	0.1393	0.0045	0.0015	0.0426	0.0366	0.0060	0.0026
Biased output	0.1030	0.1177	0.0044	0.0015	0.0395	0.0322	0.0053	0.0032
Fair output	0.1681	0.1119	0.0043	0.0010	0.0526	0.0333	0.0071	0.0032
Mixed output	0.1693	0.1292	0.0029	0.0016	0.0472	0.0386	0.0054	0.0030
Average	0.1484	0.1125	0.0038	0.0015	0.0524	0.0419	0.0064	0.0038

Table 12: Fairness Standard Deviation Across Calibration Sets: Each row shows how a specific calibration set performs with different **pruning methods**, with averages revealing the overall impact of calibration data choice on fairness metrics.

A.13 Comparative Analysis of HGLA and Wanda

We evaluated the statistical significance of differences between HGLA and Wanda by conducting paired Wilcoxon signed-rank tests across all experimental conditions. The values in Tables 14, 15, 16, and 17 represent the absolute differences using HGLA. Results marked with an asterisk (*) denote statistically significant different from Wanda ($p < 0.05$). Our analysis demonstrates that HGLA produces significant modifications across various models, pruning scenarios, and metrics, with effectiveness generally increasing at higher pruning ratios (0.3-0.4). These findings provide statistical evidence that HGLA's selective information pruning approach can reliably alter model behavior, with implications for mitigating bias and enhancing fairness in language models.

Sparsity Ratio	Calibration Sets	Llama3-8B Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Left only	-0.134	-0.160	-0.183	-0.183	-0.209	-0.165	-0.142	-0.189	-0.157	-0.152	-0.008	-0.041	0.002	-0.062	-0.011
	Right only	-0.152	-0.146	-0.123	-0.199	-0.124	-0.176	-0.165	-0.137	-0.183	-0.158	-0.035	-0.050	0.009	-0.070	-0.025
	Fair input	-0.134	-0.159	-0.227	-0.188	-0.136	-0.142	-0.175	-0.212	-0.114	-0.124	-0.076	-0.021	-0.033	-0.024	-0.027
	Mixed input	-0.095	-0.164	-0.112	-0.165	-0.131	-0.188	-0.153	-0.109	-0.179	-0.140	-0.072	-0.043	0.013	-0.059	-0.050
	Biased output	-0.132	-0.130	-0.157	-0.078	-0.206	-0.122	-0.175	-0.128	-0.174	-0.125	-0.071	-0.042	-0.052	-0.082	-0.025
	Fair output	-0.108	-0.154	-0.208	-0.166	-0.237	-0.162	-0.173	-0.138	-0.127	-0.118	-0.027	-0.028	-0.027	-0.021	-0.056
	Mixed output	-0.131	-0.142	-0.149	-0.133	-0.217	-0.173	-0.152	-0.140	-0.179	-0.166	-0.054	-0.047	-0.055	-0.061	-0.065
20%	Left only	-0.149	-0.155	-0.168	-0.162	-0.161	-0.153	-0.132	-0.138	-0.167	-0.164	-0.016	-0.015	-0.005	-0.057	0.063
	Right only	-0.136	-0.168	-0.132	-0.171	-0.095	-0.113	-0.114	-0.128	-0.130	-0.112	-0.040	-0.055	-0.045	-0.064	0.005
	Fair input	-0.142	-0.148	-0.171	-0.159	-0.137	-0.156	-0.160	-0.147	-0.137	-0.119	-0.064	-0.039	-0.057	-0.050	0.082
	Mixed input	-0.139	-0.149	-0.181	-0.152	-0.187	-0.144	-0.171	-0.150	-0.153	-0.156	-0.055	-0.033	-0.059	-0.059	-0.008
	Biased output	-0.151	-0.129	-0.176	-0.156	-0.145	-0.154	-0.159	-0.153	-0.134	-0.122	-0.059	-0.046	-0.045	-0.042	0.017
	Fair output	-0.142	-0.154	-0.182	-0.170	-0.150	-0.148	-0.132	-0.131	-0.144	-0.159	-0.016	-0.028	-0.039	-0.055	0.072
	Mixed output	-0.124	-0.170	-0.127	-0.160	-0.127	-0.141	-0.162	-0.127	-0.160	-0.152	-0.054	-0.055	-0.034	-0.060	0.017
30%	Left only	-0.151	-0.154	-0.180	-0.157	-0.072	-0.155	-0.152	-0.173	-0.167	-0.238	-0.021	-0.032	-0.031	-0.062	-0.026
	Right only	-0.165	-0.158	-0.148	-0.152	-0.034	-0.143	-0.127	-0.131	-0.143	-0.161	-0.061	-0.053	-0.021	-0.050	0.100
	Fair input	-0.143	-0.164	-0.133	-0.127	-0.020	-0.142	-0.158	-0.120	-0.154	-0.101	-0.066	-0.051	-0.066	-0.050	0.164
	Mixed input	-0.150	-0.162	-0.146	-0.128	-0.087	-0.143	-0.165	-0.162	-0.144	-0.213	-0.070	-0.045	-0.015	-0.057	-0.017
	Biased output	-0.142	-0.156	-0.161	-0.157	-0.052	-0.143	-0.170	-0.143	-0.161	-0.141	-0.079	-0.032	-0.014	-0.063	0.112
	Fair output	-0.151	-0.161	-0.150	-0.162	-0.077	-0.146	-0.158	-0.159	-0.130	-0.195	-0.053	-0.025	-0.035	-0.047	0.079
	Mixed output	-0.149	-0.166	-0.138	-0.178	-0.096	-0.143	-0.164	-0.136	-0.145	-0.194	-0.064	-0.038	-0.059	-0.071	0.064
40%	Left only	-0.154	-0.154	-0.167	-0.170	-0.047	-0.166	-0.151	-0.130	-0.157	-0.212	-0.035	-0.040	-0.049	-0.058	-0.024
	Right only	-0.146	-0.132	-0.132	-0.135	-0.008	-0.154	-0.176	-0.143	-0.156	-0.079	-0.015	-0.024	-0.033	-0.042	0.173
	Fair input	-0.134	-0.152	-0.129	-0.134	-0.178	-0.158	-0.159	-0.137	-0.139	-0.097	-0.066	-0.038	-0.056	-0.061	0.046
	Mixed input	-0.149	-0.137	-0.152	-0.140	0.006	-0.161	-0.146	-0.160	-0.158	-0.169	-0.050	-0.033	-0.011	-0.050	0.164
	Biased output	-0.155	-0.137	-0.164	-0.161	-0.086	-0.167	-0.161	-0.145	-0.165	-0.164	-0.057	-0.051	-0.038	-0.074	0.224
	Fair output	-0.147	-0.130	-0.137	-0.138	-0.086	-0.155	-0.182	-0.157	-0.170	-0.091	-0.035	-0.023	-0.055	-0.044	0.103
	Mixed output	-0.149	-0.147	-0.131	-0.140	-0.041	-0.159	-0.151	-0.128	-0.158	-0.110	-0.053	-0.046	-0.057	-0.065	0.124

(a) Political tweet summarisation

Sparsity Ratio	Calibration Sets	Llama3-8B Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	Gemma-2B Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA	TinyLlama Sparse GPT	Wanda	GBLM Pruner	GBLM Gradient	HGLA
10%	Positive only	-0.236	-0.190	-0.208	-0.220	-0.238	-0.397	-0.393	-0.401	-0.393	-0.399	-0.187	-0.145	-0.179	-0.177	-0.176
	Negative only	-0.211	-0.249	-0.266	-0.271	-0.206	-0.398	-0.416	-0.397	-0.395	-0.399	-0.172	-0.184	-0.179	-0.182	-0.177
	Fair input	-0.220	-0.257	-0.296	-0.282	-0.198	-0.404	-0.421	-0.400	-0.426	-0.406	-0.200	-0.162	-0.187	-0.182	-0.177
	Mixed input	-0.195	-0.243	-0.261	-0.261	-0.213	-0.403	-0.417	-0.411	-0.412	-0.391	-0.192	-0.171	-0.175	-0.182	-0.178
	Biased output	-0.209	-0.248	-0.253	-0.246	-0.212	-0.399	-0.391	-0.399	-0.391	-0.406	-0.201	-0.159	-0.191	-0.187	-0.171
	Fair output	-0.192	-0.253	-0.272	-0.278	-0.207	-0.402	-0.416	-0.401	-0.411	-0.392	-0.171	-0.156	-0.178	-0.173	-0.179
	Mixed output	-0.183	-0.238	-0.240	-0.222	-0.205	-0.399	-0.402	-0.385	-0.392	-0.395	-0.190	-0.164	-0.175	-0.191	-0.180
20%	Positive only	-0.196	-0.239	-0.231	-0.237	-0.200	-0.408	-0.392	-0.414	-0.415	-0.406	-0.165	-0.150	-0.190	-0.183	-0.153
	Negative only	-0.219	-0.248	-0.262	-0.246	-0.176	-0.409	-0.408	-0.398	-0.402	-0.409	-0.172	-0.173	-0.160	-0.176	-0.151
	Fair input	-0.186	-0.205	-0.218	-0.222	-0.197	-0.401	-0.383	-0.396	-0.375	-0.418	-0.176	-0.173	-0.161	-0.159	-0.138
	Mixed input	-0.203	-0.255	-0.246	-0.242	-0.190	-0.407	-0.397	-0.403	-0.411	-0.397	-0.170	-0.171	-0.155	-0.176	-0.190
	Biased output	-0.202	-0.225	-0.231	-0.228	-0.212	-0.395	-0.387	-0.392	-0.406	-0.398	-0.186	-0.162	-0.160	-0.175	-0.163
	Fair output	-0.174	-0.248	-0.244	-0.240	-0.227	-0.404	-0.392	-0.407	-0.401	-0.400	-0.181	-0.177	-0.165	-0.179	-0.151
	Mixed output	-0.206	-0.215	-0.207	-0.265	-0.208	-0.399	-0.402	-0.391	-0.401	-0.398	-0.178	-0.154	-0.171	-0.162	-0.176
30%	Positive only	-0.193	-0.224	-0.214	-0.206	-0.210	-0.404	-0.393	-0.402	-0.393	-0.408	-0.182	-0.158	-0.166	-0.170	-0.159
	Negative only	-0.183	-0.210	-0.187	-0.213	-0.191	-0.410	-0.405	-0.418	-0.400	-0.394	-0.162	-0.164	-0.154	-0.158	-0.141
	Fair input	-0.210	-0.252	-0.253	-0.268	-0.190	-0.416	-0.404	-0.410	-0.407	-0.411	-0.182	-0.151	-0.171	-0.163	-0.136
	Mixed input	-0.199	-0.229	-0.231	-0.228	-0.218	-0.410	-0.392	-0.405	-0.415	-0.427	-0.196	-0.175	-0.180	-0.191	-0.151
	Biased output	-0.185	-0.250	-0.237	-0.248	-0.234	-0.418	-0.405	-0.411	-0.434	-0.378	-0.181	-0.156	-0.167	-0.183	-0.165
	Fair output	-0.178	-0.231	-0.230	-0.252	-0.241	-0.412	-0.408	-0.417	-0.424	-0.402	-0.171	-0.159	-0.180	-0.176	-0.137
	Mixed output	-0.182	-0.227	-0.236	-0.251	-0.224	-0.408	-0.401	-0.404	-0.402	-0.431	-0.196	-0.153	-0.165	-0.162	-0.173
40%	Positive only	-0.200	-0.236	-0.222	-0.220	-0.268	-0.417	-0.378	-0.385	-0.380	-0.323	-0.193	-0.160	-0.171	-0.182	-0.161
	Negative only	-0.178	-0.248	-0.262	-0.263	-0.304	-0.406	-0.378	-0.383	-0.359	-0.283	-0.173	-0.160	-0.168	-0.162	-0.224
	Fair input	-0.196	-0.261	-0.269	-0.280	-0.262	-0.419	-0.392	-0.391	-0.385	-0.356	-0.176	-0.165	-0.159	-0.166	-0.192
	Mixed input	-0.199	-0.289	-0.289	-0.304	-0.385	-0.413	-0.379	-0.371	-0.382	-0.372	-0.185	-0.155	-0.160	-0.143	-0.171
	Biased output	-0.182	-0.262	-0.282	-0.284	-0.307	-0.421	-0.378	-0.361	-0.358	-0.348	-0.178	-0.144	-0.169	-0.136	-0.196
	Fair output	-0.206	-0.270	-0.322	-0.302	-0.290	-0.424	-0.396	-0.371	-0.379	-0.310	-0.186	-0.165	-0.207	-0.120	-0.196
	Mixed output	-0.218	-0.272	-0.344	-0.315	-0.377	-0.436	-0.381	-0.313	-0.380	-0.347	-0.174	-0.140	-0.173	-0.082	-0.183

(b) Review summarisation

Table 13: Raw Second-Order SPD

Ratio	Llama3-8B			Gemma-2B			TinyLlama		
	SPD	SOF	UER	SPD	SOF	UER	SPD	SOF	UER
0.1	-0.060	-0.006	0.006	-0.046	0.005	0.002	0.038	-0.007	0.011
0.2	-0.004	-0.004	0.000	0.045	0.010	-0.002	0.077	0.003	0.010
0.3	0.119	-0.005	-0.001	-0.052	0.004*	-0.013	-0.111*	0.004*	0.024
0.4	0.170*	0.003	0.015	0.112	0.015	0.004	-0.258*	0.010	0.011

Table 14: Political tweet summarisation—Right Information Only Pruned (HGLA vs Wanda)

Ratio	Llama3-8B			Gemma-2B			TinyLlama		
	SPD	SOF	UER	SPD	SOF	UER	SPD	SOF	UER
0.1	-0.232	-0.003*	-0.001	-0.035	0.005	0.001	0.065	-0.012*	0.015*
0.2	-0.134	0.004	0.002	-0.059*	0.015	0.000	-0.037*	-0.003*	0.015
0.3	0.043	-0.002	0.002	-0.205*	0.009	-0.008	0.037	-0.007	0.010
0.4	0.092*	0.001*	0.019	-0.153	0.017	0.002	0.040	-0.004	0.010

Table 15: Political tweet summarisation—Left Information Only Pruned (HGLA vs Wanda)

Ratio	Llama3-8B			Gemma-2B			TinyLlama		
	SPD	SOF	UER	SPD	SOF	UER	SPD	SOF	UER
0.1	0.020*	0.000	-0.003	-0.013	-0.001	0.001	0.031	-0.001*	0.003*
0.2	0.097	-0.003	0.004	-0.028*	0.002	0.000	0.077	-0.001*	0.002
0.3	0.077*	-0.003	0.000	-0.032	-0.001*	0.005	0.064	-0.001	0.005
0.4	-0.040	0.001*	0.003*	0.140*	-0.003*	0.012	0.061	-0.001	-0.001

Table 16: Review summarisation—Positive Information Only Pruned (HGLA vs Wanda)

Ratio	Llama3-8B			Gemma-2B			TinyLlama		
	SPD	SOF	UER	SPD	SOF	UER	SPD	SOF	UER
0.1	0.085	-0.003	-0.001	-0.013	0.001	0.000	0.029	-0.002*	0.003*
0.2	0.145	-0.003	0.000	-0.034*	0.001	0.000	0.080	-0.001	0.000
0.3	0.114*	0.001*	0.001*	-0.004	-0.003*	0.005*	0.101	-0.001	0.007
0.4	-0.112	-0.001	0.001	0.219*	-0.001*	0.007*	-0.065*	0.000	0.000

Table 17: Review summarisation—Negative Information Only Pruned (HGLA vs Wanda)