Coverage-based Fairness in Multi-document Summarization

Anonymous ACL submission

Abstract

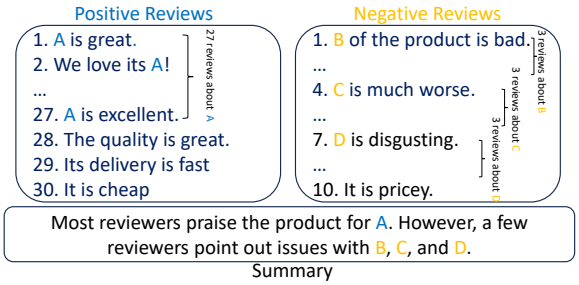
Fairness in multi-document summarization (MDS) measures whether a system can generate a summary fairly representing information from documents with different social attribute values. Fairness in MDS is crucial since a fair summary can offer readers a comprehensive view. Previous works focus on quantifying summary-level fairness using Proportional Representation, a fairness measure based on Statistical Parity. However, Proportional Representation does not consider redundancy in input documents and overlooks corpus-level unfairness. In this work, we propose a new summary-level fairness measure, **Equal Coverage**, which is based on coverage of documents with different social attribute values and considers the redundancy within documents. To detect the corpus-level unfairness, we propose a new corpus-level measure, **Coverage Parity**. Our human evaluations show that our measures align with the human perception of fairness. Using our measures, we evaluate the fairness of ten different LLMs. We find that Llama2 is the fairest among all evaluated LLMs. We also find that almost all LLMs overrepresent different social attribute values.

1 Introduction

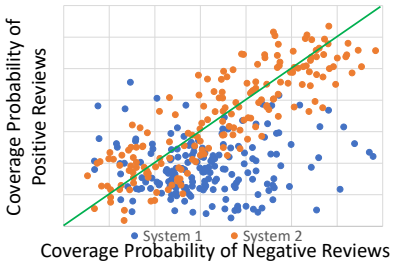
Multi-document summarization (MDS) systems summarize the salient information from multiple documents about an entity, such as news articles about an event or reviews of a product. Typically, such documents are associated with *social attributes* e.g. political ideology in news and sentiments in reviews. Documents with different social attributes tend to have diverse information or conflicting opinions. A summary for them should fairly represent differing opinions across documents.

Fairness in MDS measures whether a system can generate summaries fairly representing information from documents with different social attribute values. It can be measured at a *summary-level*– quantifying how fairly an individual summary represents documents with different social attribute values or at a *corpus-level*–quantifying how fairly a corpus of summaries as a whole represents documents with different social attribute values. Previous works in this area measured fairness in extractive or abstractive settings (Shandilya et al., 2018; Olabisi et al., 2022; Zhang et al., 2023; Huang et al., 2024). These works generally evaluate the fairness of a system as the aggregated summary-level fairness of its generated summaries. It is measured using Proportional Representation–a fairness measure based on Statistical Parity (Verma and Rubin, 2018). It requires that a document sentence being selected as a summary sentence is independent of its originating document's social attribute value.

The definition of Proportional Representation suffers from two key problems. The first problem is that Proportional Representation does not consider the redundancy in input documents, common in multi-document settings. For example, in review summarization, suppose 75% of reviews are positive and most of them discuss topic A, while 25% of reviews are negative and evenly discuss topics B, C, and D (see Fig. 1a). According to the definition of Proportional Representation, if the system selects information from input independent of the social attribute value, a fair summary should ideally have 75% of information from positive reviews and 25% percent of information from negative reviews. The summary shown in the figure will be considered unfair according to this definition because, unlike the input, it contains more negative information than positive information. However, considering both redundancy and input sentiment distribution, this summary is fair since it covers equal proportions of reviews for both sentiments while avoiding redundant information. To address this, we propose a new summary-level coverage-based fairness measure, **Equal Coverage**. Unlike Proportional Representation, Equal Coverage requires a docu-



(a) Existing summary-level fairness measure, Proportional Representation, does not consider redundancy common in MDS. It incorrectly considers the illustrated summary unfair since, unlike the input, it contains more negative (but diverse) information than positive (but redundant) information. Our proposed measure, Equal Coverage, correctly considers this summary fair because it covers equal proportions of negative and positive reviews.



(b) Existing summary-level fairness measures can overlook corpus-level unfairness. Each point in this figure represents a summary sample. System 2 is fairer than System 1 since it has equal chances of overrepresenting negative (below the green line) and positive (above the green line) reviews while System 1 tends to overrepresent negative reviews. Our proposed measure, Coverage Parity, can correctly identify System 2 as fairer than System 1.

Figure 1: Issues with existing fairness measures for summary-level (a) and corpus-level (b) fairness.

ment being *covered* by a summary sentence to be independent of its social attribute value. Since a summary sentence can cover multiple documents with similar contents, Equal Coverage can address redundancy common for MDS.

The second problem is evaluating the fairness of an MDS system only using summary-level fairness measures can overlook corpus-level unfairness. Consider System 1 in Fig. 1b. Most of its summaries (blue dots) have a higher coverage probability for negative reviews than positive reviews. We observe that System 1 is unfair because its summaries *overrepresent* negative reviews. System 2 is fairer because its summaries (orange dots) have an equal chance of overrepresenting negative or positive reviews. Since individual summaries from both systems overrepresent negative or positive reviews, their summary-level fairness scores may be comparable. Hence, aggregated summary-level fairness scores cannot identify that System 2 is collectively fairer than System 1. To address this problem, we propose a new corpus-level fairness metric, **Coverage Parity**. Motivated by Best-Worst Scaling (Louviere et al., 2015), Coverage Parity is based on the principle that different social attribute values should have equal chances of being most overrepresented or underrepresented in summaries. Therefore, it can check whether the systems are equally (un)fair on different social attributes and identify which social attribute is overrepresented or underrepresented.

Our human evaluation shows that our measures align more with the human perception of fairness than Proportional Representation. Using our measures, we then evaluate the fairness of ten different Large Language Models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023) in diverse domains: news, tweets, and reviews. For these domains, we consider social attributes having significant real-world impacts: ideologies and stances for news and tweets and sentiment for reviews. Our experiments find that for Llama2 and Claude3, larger models are fairer than smaller models, but this trend is inconsistent for GPTs and Mixtral. Our experiments also find that almost all LLMs tend to overrepresent certain social attribute values in each domain. It is an important finding that users can use to calibrate their perception before using LLM-generated summaries. It can also be used by developers to build fairer LLM-based summarization systems.

To conclude, our contributions are three-fold:

- We propose a new summary-level fairness measure, Equal Coverage, which incorporates redundancy of input information, common in MDS;
- We propose a new corpus-level fairness measure, Coverage Parity to detect corpus-level unfairness;
- We evaluate the fairness of LLMs using these two measures in various domains.

2 Related Work

Shandilya et al. (2018, 2020); Dash et al. (2019) propose to measure the summary-level fairness in MDS under the extractive setting using Proportional Representation. They propose an in-processing method to improve the fairness of extractive summaries. Similarly, Keswani and Celis (2021) uses Proportional Representation to measure the fairness of summaries under different dis-

tributions of social attributes in input documents and proposes a post-processing method to improve fairness. Olabisi et al. (2022) uses the same measure to measure fairness and proposes a clustering-based pre-processing method to improve fairness.

Recently, Zhang et al. (2023); Huang et al. (2024) extend Proportional Representation to measure fairness under the abstractive setting. To estimate the distribution of social attributes in a summary, Zhang et al. (2023) maps the summary back to the originating documents, while Huang et al. (2024) uses a finetuned model to estimate the social attribute of each summary sentence. Lei et al. (2024) measures the fairness similarly as Huang et al. (2024) and proposes to improve the fairness of abstractive summaries using reinforcement learning with a polarity distance reward. However, these works generally measure fairness using Proportional Representation, which has the limitations discussed in the introduction. In our experiments, we use Proportional Representation as a baseline.

3 Notation

We use G to denote all samples for evaluating the fairness of a system on a social attribute. Each sample $(D, S) \in G$ contains a document set $D = \{d_1, ..., d_n\}$ and a summary S generated by the system for these documents, where d_i denotes the i-th document. Each input document d_i is labeled with an social attribute value $a_i \in \{1, ..., K\}$.

4 Fairness Measures

In this section, we describe our proposed measures Equal Coverage (Sec. 4.1) and Coverage Parity (Sec. 4.2).

4.1 Equal Coverage

Equal Coverage is a summary-level fairness measure for measuring the fairness of a summary, S, for document sets, D. Equal Coverage is based on the principle that the documents with different social attribute values should have equal probabilities for being covered by a summary sentence. We denote the probability that a random document $d \in D$ whose social attribute value a is k is covered by a random summary sentence $s \in S$ as $p(d, s|a = k)$. This is referred to as the coverage probability for the social attribute value k. Similarly, we denote the probability that a random document, d, is covered by a random summary sentence $s \in S$ as $p(d, s)$, which is referred to as the coverage probability for a document. For a fair summary S according to Equal Coverage, coverage of a random document d should be independent of its social attribute value:

$$p(d, s|a = k) = p(d, s), \ \forall k. \quad (1)$$

However, two issues arise with summaries' complex sentence structures. First, a summary sentence can combine information from several documents, making it difficult to attribute the sentence to any single document. Second, sentence lengths vary significantly based on social attribute values. For example, summary sentences about positive sentiment can be long, such as 'the delivery is fast and it is worth the price.'. In contrast, those about negative sentiment can be much shorter, such as 'delivery is too slow.'. The length difference can skew the coverage probability for different social attribute values $p(d, s|a = i)$. To address these issues, Equal Coverage splits the summary sentences $s \in S$ into multiple simpler sentences by prompting GPT-3.5 motivated by Bhaskar et al. (2023); Min et al. (2023). For example, compound sentences are split into simple sentences that describe a single fact. We denote the j-th summary sentence after split as s_j. Further details are in App. A.1.

Equal Coverage estimates the coverage probability for different social attribute values $p(d, s|a = k)$ and for a document $p(d, s)$ based on the probability $p(d_i, s_j)$ that a document d_i is covered by a summary sentence s_j. The probability $p(d_j, s_k)$ is estimated as the entailment probability that the document d_j entails the summary sentence s_k by a textual entailment model (Williams et al., 2018). Motivated by Laban et al. (2022), Equal Coverage divides documents into chunks of M words. The l-th chunk of the document d_j is denoted as $d_{j,l}$. The entailment model estimates the probability $p(d_j, s_k)$ as the maximum entailment probability $p(d_{j,l}, s_k)$ between the chunk $d_{j,l}$ and the summary sentence s_k:

$$p(d_i, s_j) = \max\{p(d_{i,l}, s_j)|d_{i,l} \in d_i\}, \quad (2)$$

where $p(d_{i,l}, s_j)$ is the probability that the document chunk $d_{i,l}$ entails the summary sentence s_j as per the entailment model. Based on the probability $p(d_i, s_j)$ that the document d_i is covered by the summary sentence s_j, the coverage probability for social attribute value i, $p(d, s|a = k)$ is empirically estimated as:

$$p(d, s|a = k) = \frac{1}{|D_k||S|} \sum_{d_i \in D_k} \sum_{s_j \in S} p(d_i, s_j), \quad (3)$$

where D_k is the set of documents d_i whose social attribute value a_i is k. Similarly, the coverage probability for a document, $p(d, s)$ is estimated as:

$$p(d, s) = \frac{1}{|D||S|} \sum_{d_j \in D} \sum_{s_k \in S} p(d_j, s_k). \quad (4)$$

Recall that for a fair summary S according to Equal Coverage, coverage probabilities for different social attribute values $p(d, s|a = k)$ should equal the coverage probability for a document $p(d, s)$. Therefore, Equal Coverage measures the summary-level fairness $EC(D, S)$ as:

$$EC(D, S) = \frac{1}{K} \sum_{k=1}^{K} |p(d, s) - p(d, s|a = k)|. \quad (5)$$

A high Equal Coverage value $EC(D, S)$ indicates less fairness because there are big differences between coverage probabilities for different social attribute values $p(d, s|a = k)$.

To determine if the differences between coverage probabilities for different social attribute values $p(d, s|a = k)$ are statistically significant, we perform a permutation test (Pitman, 1937). We randomly permute social attribute values a_i of documents and use them to calculate the permuted Equal Coverage value $\hat{EC}(D, S)$ 5000 times. The statistical significance of the difference is estimated as the proportion of permuted Equal Coverage values $\hat{EC}(D, S)$ greater than itself. We denote the proportion of samples (D, S) where the difference between coverage probabilites for different social attribute values is statistically significant ($p < 0.05$), $R_{EC}(G)$, as the proportion of unfair summaries, which is used to evaluate the fairness of the system.

4.2 Coverage Parity

Coverage Parity is a corpus-level fairness measure designed to measure the fairness of a system of all samples G. Motivated by Best-Worst Scaling (Louviere et al., 2015), Coverage Parity is based on the principle that different social attribute values should have equal chances of being the most overrepresented or underrepresented in summaries. It only considers the most overrepresented and underrepresented social attribute values because different samples' document sets D have different combinations of social attribute values. For instance, a document set can contain only positive and neutral reviews, while another can contain reviews with all sentiments. Therefore, Coverage Parity ensures that each sample contributes equally to it.

For a fair MDS system, Coverage Parity requires that the average coverage probability difference $c_k(d, s) = p(d, s) - p(d, s|a = k)$ is close to zero when the social attribute value k is the most overrepresented or underrepresented among all samples. In a sample (D, S), we define the social attribute value i as the most overrepresented if its coverage probability difference $c_k(d, s)$ is the maximum $\max_k\{c_k(d, s)\}$ among all social attribute values $k \in \{1, ..., K\}$, and the most underrepresented if its coverage probability difference $c_k(d, s)$ is the minimum $\min_k\{c_k(d, s)\}$. We collect these coverage probabilities differences $c_k(d, s)$ from all samples where social attribute value k is the most overrepresented or underrepresented in a set C_k.

Based on the average of the set of coverage probability differences C_i, Coverage Parity measures the fairness of the MDS system:

$$CP(G) = \frac{1}{K} \sum_{i=1}^{K} |\mathbb{E}(C_k)|. \quad (6)$$

A high Coverage Parity value $CP(G)$ indicates less fairness since it suggests that the chances of being overrepresented or underrepresented are very different for some social attribute values. Based on whether the average coverage probability differences, $\mathbb{E}(C_k)$, is greater or less than zero, we can also identify which social attribute value tends to be overrepresented or underrepresented.

5 Experimental Setup

We now describe experiments to evaluate the fairness of LLMs using our measures.

5.1 Datasets

We conduct experiments on five different datasets from the three domains: reviews, tweets, and news. These datasets are the Amazon (Ni et al., 2019), MITweet (Liu et al., 2023), SemEVal (Mohammad et al., 2016), the Article Bias (Baly et al., 2020), and the News Stance (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017; Hanselowski et al., 2019) datasets. Tab. 1 shows the statistics of these datasets along with their social attribute values.

We observe that for some datasets, the fairness of summaries depends on the distributions of social attribute values in the input documents (Sec. 5.7).To balance social attribute values' impacts on fairness, we perform stratified sampling to collect 300 input document sets, D, for each dataset. Among these sampled sets, input document sets D dominated by

	Domain	Social Attribute	Social Attribute Values	Input Doc. Set Size	Avg. Doc. Len.
Amazon	Review	Sentiment	{negative, neutral, positive}	8	40
MITweet	Tweet	Political Ideology	{left, center, right}	20	34
Article Bias	News	Political Ideology	{left, center, right}	4-8	436
SemEval	Tweet	Stance toward Target	{support, against}	30	17
News Stance	News	Stance toward Target	{support, against}	4-8	240

Table 1: Dataset statistics for fairness evaluation in MDS

different social attribute values have equal proportions. The stratified sampling does not affect the calculation of our fairness measures. More details of preprocessing are in App. A.3.

5.2 Implementation Details

We evaluate the fairness of five families of LLMs: GPTs (Ouyang et al., 2022; Achiam et al., 2023) (GPT-3.5-0124, GPT-4-turbo-2024-04-09), Llama2 (Touvron et al., 2023) (Llama2-7b, Llama2-13b, Llama2-70b), Mixtral (Jiang et al., 2023, 2024) (Mistral-7B-Instruct-v0.1, Mixtral-8x7B-Instruct-v0.1), Gemma (Team et al., 2024) (gemma-1.1-7b-it), and Claude3 (Bai et al., 2022) (claude-3-haiku-20240307, claude-3-sonnet-20240229). For comparison, we also evaluate the fairness of COOP (Iso et al., 2021) on the Amazon dataset and the fairness of PEGASUS (Zhang et al., 2020) and PRIMERA (Xiao et al., 2022) finetuned on the Multi-News (Fabbri et al., 2019) on the other datasets. The summary length is 100 words for the Article Bias and News Stance datasets and 50 for the other datasets. We describe summarization prompts in App. A.5.

To estimate the probability that a document is covered by a summary sentence (Eqn. 2), we use RoBERTa-large finetuned on the MNLI dataset (Williams et al., 2018). However, we show that our measures are independent of this choice (Sec. 5.3). To estimate the entailment probability using RoBERTa-large, we divide documents into chunks of $W = 100$ words. The chunk size W is tuned to maximize the proportion of summary sentences whose originating documents can be identified by RoBERTa-large (App. A.2) since LLMs are less prone to factual errors (Goyal et al., 2022).

5.3 Choice of Entailment Models

The implementations of our measures are independent of the choice of entailment model. To demonstrate this, we calculate our measures using three different textual entailment models: RoBERTa finetuned on the MNLI dataset; DeBERTa-large finetuned on multiple entailment datasets (Laurer et al., 2024); and ALBERT-xl (Lan et al., 2019) finetuned on the MNLI and VitaminC (Schuster et al., 2021) datasets. We report the Spearman correlations between the Equal Coverage values, $EC(D, S)$, of each summary generated by all LLMs, obtained using these entailment models. We also report the Spearman correlations between the Coverage Parity value, $CP(G)$, of each LLM. The results are in Table 2. From the table, we can observe strong correlations between measures obtained using different textual entailment models on most datasets. It shows that these measures are independent of the choice of entailment models.

	Ro. vs De.	Ro. vs Al.	De. vs Al.
	Equal Coverage		
Amazon	0.544	0.439	0.358
MITweet	0.430	0.442	0.357
Article Bias	0.470	0.709	0.429
SemEval	0.426	0.370	0.306
News Stance	0.712	0.780	0.703
	Coverage Parity		
Amazon	0.867	0.939	0.733
MITweet	-0.127	0.006	0.673
Article Bias	0.624	0.915	0.612
SemEval	0.612	0.891	0.539
News Stance	0.903	0.915	0.939

Table 2: Spearman correlations between Equal Coverage (top) values and Coverage Parity (bottom) values computed using different entailment models: RoBERTa (Ro.), Deberta (De.), and ALBERT (Al.). Correlations that are statistically significant are underlined. We observe strong correlations, indicating that our measures are independent of the choice of the entailment model.

5.4 Human Perception of Fairness

We perform a human evaluation to determine which measure, Equal Coverage or Proportional Representation, aligns more with human perception of fairness. For Proportional Representation, we use the BARTScore (Yuan et al., 2021) implementation proposed by Zhang et al. (2023), as it shows the highest correlation with human perception.

Evaluating human perception of fairness is challenging due to the need to review entire input document sets. Therefore, we perform experiments on the Amazon dataset which only contains eight short reviews per input document set (Tab. 1). To

simplify the evaluation, we only consider input document sets containing only negative and positive but not neutral reviews, displayed in two randomized columns. For each input document set, we consider the summary generated by GPT-3.5 since it shows medium-level fairness (Tab. 3, 4). To further simplify the evaluation, we focus on summaries where Equal Coverage and Proportional Representation disagree on their fairness. We randomly select 25 samples containing input document sets and corresponding summaries that meet these criteria. Each sample is annotated by three annotators recruited from Amazon Mechanical Turk. The annotators should be from English-speaking countries and have HIT Approval Rates greater than 98%. More details are in App. A.4.

For each sample, annotators are asked to identify all unique negative and positive opinions in the input document set. They then evaluate whether the summary reflects these opinions and classify the summary as leaning negative, fair, or leaning positive. Out of these 25 samples, human perception of fairness aligns more with Equal Coverage in 17 samples, while aligns more with Proportional Representation in 8 samples. The difference is statistically significant ($p < 0.05$) using paired bootstrap resampling (Koehn, 2004). It shows that Equal Coverage aligns more with human perception of fairness than Proportional Representation.

We compare Coverage Parity with second-order fairness based on Proportional Representation to evaluate their effectiveness in identifying overrepresented sentiments on the group level. For this, we perform bootstrapping on the 25 samples to generate 5000 groups of 25 bootstrap samples each. For each groups of bootstrap samples, we define the sentiment that most bootstrap samples' summary leaning toward based on human annotation as the sentiment overrepresented on the group level. We then compare the overrepresented sentiment based on human annotation with those identified by the Coverage Parity and second-order fairness. We find that Coverage Parity aligns with human annotation in 96% of the groups, while second-order fairness aligns in 4% of the groups. It shows that Coverage Parity aligns more with human perception of fairness than the second-order fairness.

5.5 Summary-level Fairness Evaluation

To evaluate the summary-level fairness of different LLMs, we report the proportion of unfair summaries according to Equal Coverage, $R_{EC}(G)$. We also report an *Overall* score which is the average of normalized $R_{EC}(G)$ ($[0, 1]$) on all datasets. The results are in Tab. 3.

	Amazon	MITweet	Article Bias	SemEval	News Stance	Overall
GPT-3.5	0.127	0.050	0.210	0.103	0.433	0.517
GPT-4	0.103	0.067	0.187	0.147	0.380	0.469
Llama2-7b	0.127	0.090	0.236	**0.083**	0.427	0.686
Llama2-13b	0.107	0.080	0.239	0.087	0.315	0.441
Llama2-70b	**0.090**	0.057	0.177	0.137	**0.308**	0.224
Mistral-7b	0.127	0.082	0.193	0.115	0.334	0.526
Mixtral-8x7b	0.117	0.097	**0.172**	0.128	0.372	0.549
Gemma	0.101	**0.047**	0.200	0.090	0.407	0.264
Claude3-haiku	0.110	0.063	0.174	0.093	0.500	0.411
Claude3-sonnet	0.107	0.070	0.174	0.120	0.401	0.401
COOP	0.204	-	-	-	-	-
PEGASUS	-	0.104	0.158	0.199	0.390	-
PRIMERA	-	0.118	0.134	0.118	0.377	-

Table 3: Proportion of unfair summaries and overall scores on different datasets according to Equal Coverage. A lower value indicates a fairer system. **Bold** indicates the fairest system.

From the table, we observe Llama2-70b is the fairest. Among smaller LLMs (with 7-billion parameters), Gemma is the fairest. We also observe that almost all evaluated LLMs are fairer than COOP on the Amazon dataset, and PEGASUS and PRIMERA on the MITweet and Article Bias datasets. For the comparison within families of GPTs, Llama2, Claude3, we observe that larger models are generally fairer. However, Mixtral-8x7b is less fair than Mistral-7b although Mixtral-8x7b is larger, which might be because it scales its size using the mixture of smaller models.

Previous works by Zhang et al. (2023) and Lei et al. (2024) evaluate the fairness of LLMs using Proportional Representation. When evaluating the fairness on sentiments in the review domain, our results using Equal Coverage are consistent with these works in the finding that GPT-4 is fairer than GPT-3.5, and both are fairer than COOP. However, they find that Llama2-13b is the fairest, while our results show that larger models are fairer for Llama2. When evaluating the fairness on political ideologies in the tweet domain, our results are consistent with these works on that GPT-4 is fairer than GPT-3.5. However, they find that smaller models are fairer for Llama2, and GPTs are less fair than Llama2 models. Contrarily, we show that larger models are fairer for Llama2, and GPTs are fairer than Llama2-7b. As shown in Sec. 5.4, Equal Coverage aligns more with human perception, suggesting our results better reflect human judgments.

5.6 Corpus-level Fairness Evaluation

To evaluate the corpus-level fairness of different LLMs, we report the Coverage Parity, $CP(G)$, on

	Amazon			MITweet			Article Bias			SemEval			News Stance			Overall
	$CP(G)$	over	under	$CP(G)$	over	under	$CP(G)$	over	under	$CP(G)$	over	under	$CP(G)$	over	under	
GPT-3.5	0.032	neg.	pos.	0.012	left	center	**0.015**	center	left	0.012	sup.	aga.	0.018	aga.	sup.	0.475
GPT-4	0.029	neg.	pos.	0.011	left	center	0.024	right	left	0.011	sup.	aga.	0.038	aga.	sup.	0.647
Llama2-7b	0.037	neg.	pos.	0.013	left	center	0.026	right	left	0.010	sup.	aga.	0.018	aga.	sup.	0.597
Llama2-13b	0.017	neg.	pos.	0.009	left	center	0.025	right	left	0.014	sup.	aga.	0.022	aga.	sup.	0.460
Llama2-70b	**0.008**	neg.	neu.	0.008	left	center	0.025	right	left	0.007	sup.	aga.	**0.012**	aga.	sup.	**0.177**
Mistral-7b	0.024	neg.	pos.	0.009	left	center	0.029	right	left	0.014	sup.	aga.	0.023	aga.	sup.	0.551
Mixtral-8x7b	0.017	neg.	pos.	0.010	left	center	0.036	right	left	0.003	sup.	aga.	0.020	aga.	sup.	0.394
Gemma	0.025	neg.	pos.	0.015	left	center	0.020	right	left	0.014	sup.	aga.	0.032	aga.	sup.	0.718
Claude3-haiku	0.028	neg.	pos.	0.013	left	center	0.016	center	left	0.012	sup.	aga.	0.042	aga.	sup.	0.651
Claude3-sonnet	0.025	neg.	pos.	**0.008**	left	center	0.018	**right**	left	**0.002**	sup.	aga.	0.030	aga.	sup.	0.264
COOP	0.068	pos.	neg.	-	-	-	-	-	-	-	-	-	-	-	-	-
PEGASUS	-	-	-	0.009	right	left	0.035	right	left	0.001	sup.	aga.	0.022	sup.	aga.	-
PRIMERA	-	-	-	0.014	right	left	0.032	right	left	0.002	aga.	sup.	0.023	sup.	aga.	-

Table 4: Coverage Parity, $CP(G)$, and the most overrepresented (over) and underrepresented (under) social attribute values, and overall scores on different datasets. A lower value of $CP(G)$ indicates a fairer system. **Bold** indicates the fairest system. The social attribute values whose average coverage probability differences are statistically significantly (p<0.05) different from zero based on Bootstrapping are underlined.

different datasets. For each dataset, we report the most overrepresented and underrepresented social attribute value i whose average coverage probability difference, $\mathbb{E}(C_i)$, is the maximum and minimum respectively. We also report an *Overall* score which is the average of normalized $CP(G)$ ($[0, 1]$) on all datasets. The results are in Tab. 4.

From the table, we observe that Llama2-70b is the fairest. Among smaller LLMs (with 7-billion parameters), Mistral-7b is the fairest. We also observe that evaluated LLMs are less fair than PEGASUS and PRIMERA on the SemEval dataset. While comparing within each family of LLMs, we observe that larger models are fairer for the families of Llama2, Mixtral, and Claude3. However, for the family of GPTs, GPT-4 is less fair than GPT-3.5, suggesting that larger models are not necessarily more fair on the system level. Besides, we observe that the fairness measured by Coverage Parity and Equal Coverage are different on some datasets. The difference indicates that we should consider both summary-level fairness and corpus-level fairness for comprehensively measuring fairness in MDS.

We can also observe that most LLMs overrepresent and underrepresent certain social attribute values on different datasets. For the Amazon dataset, most LLMs overrepresent negative reviews and underrepresent positive reviews. Contrarily, COOP overrepresents positive reviews and underrepresents negative reviews. For the MITweet and Article Bias datasets, all LLMs overrepresent left-leaning documents in the tweet domain, while most LLMs overrepresent right-leaning documents in the news domain. Contrarily, PEGASUS and PRIMERA overrepresent right-leaning documents for all domains. We can observe the same pattern for the SemEval and News Stance datasets. All LLMs overrepresent documents supporting the target in the tweet domain but overrepresent documents against the target in the news domain. These results indicate that summaries generated by LLMs might overrepresent documents with certain social attribute values. Users should know this before they make judgments based on these summaries. For example, users should know that a review summary generated by LLMs for a product might unfairly overrepresent negative reviews so they can calibrate their perception of the product accordingly. Developers can also build fairer LLMs for MDS based on these results.

5.7 Fainess under Different Distributions of Social Attributes

We perform experiments to evaluate whether the fairness of LLMs changes under different distributions of social attribute values in input document sets. For this, we divide all samples G into K non-overlapping sets: $\{G_1, ...G_K\}$ based on distributions of social attribute values. Each set G_i includes samples where most documents $d \in D$ have a social attribute value of i. We denote the set G_i as the set dominated by social attribute value i. To measure differences of the summary-level fairness measured by Equal Coverage under different distributions, we use maximum differences of proportions of unfair summaries $R_{EC}(G_i)$ (Sec. 4.1) on sets dominated by different social attribute values G_i. To measure differences of the corpus-level fairness measured by Coverage Parity under different distributions, we use maximum differences of average coverage probability differences for the same social attribute value $\mathbb{E}(C_j)$ (Sec. 4.2) on

	Amazon	MITweet	Article Bias	SemEval	News Stance
			Equal Coverage		
GPT-3.5	0.080	0.050	0.070	0.067	0.061
GPT-4	0.070	0.040	0.040	0.067	0.044
Llama2-7b	0.110	0.040	0.030	0.020	0.032
Llama2-13b	0.010	0.020	0.060	0.053	0.022
Llama2-70b	0.100	0.060	0.122	0.052	0.028
Mistral-7b	0.070	0.071	0.051	0.103	0.031
Mixtral-8x7b	0.120	0.039	0.051	0.053	0.025
Gemma	0.069	0.050	0.070	0.067	0.038
Claude3-haiku	0.060	0.040	0.042	0.023	0.053
Claude3-sonnet	0.110	0.050	0.060	0.028	0.107
			Coverage Parity		
GPT-3.5	0.085	0.030	0.029	0.011	0.093
GPT-4	0.086	0.030	0.075	0.009	0.078
Llama2-7b	0.128	0.027	0.038	0.011	0.091
Llama2-13b	0.088	0.016	0.074	0.014	0.116
Llama2-70b	0.099	0.023	0.044	0.008	0.068
Mistral-7b	0.072	0.012	0.046	0.017	0.120
Mixtral-8x7b	0.069	0.022	0.027	0.008	0.061
Gemma	0.089	0.029	0.036	0.017	0.063
Claude3-haiku	0.103	0.021	0.061	0.013	0.064
Claude3-sonnet	0.083	0.022	0.072	0.010	0.093

Table 5: Differences of Equal Coverage and Coverage Parity under different distributions of social attribute values in input document sets. Statistically significant ($p<0.05$) differences based on Bootstrapping are underlined. We observe significant differences in Coverage Parity for most LLMs. It indicates that the social attribute values overrepresented or underrepresented change significantly under different distributions.

different sets G_i. The results are in Table 5.

From the table, we observe that Equal Coverage remains stable under different distributions. However, for most LLMs, the differences of Coverage Parity are significant. It suggests that the social attribute values overrepresented or underrepresented change significantly under different distributions of social attribute values in input document sets. We further analyze the data from the Amazon and News Stance datasets, where the differences are significant for all LLMs. On these two datasets, almost all LLMs tend to overrepresent the social attribute values that dominate the input document sets. Therefore, corpus-level fairness measures like Coverage Parity might mistakenly judge which social attribute values are overrepresented based on the size of set dominated by different social attribute values $|G_i|$, showing the need of balanced datasets for evaluating the fairness. We address this issue through stratified sampling (Sec. 5.1).

5.8 LLM Perception of Fairness

We perform experiments to evaluate which measure, Equal Coverage or Proportional Representation, aligns more with LLMs' perception of fairness. This is an exploratory experiment, and we do not assume the LLMs' perception of fairness as ground truth. We prompt an LLM to generate a summary for an input document set, then prompt it again to generate a fair summary for the same set. The second prompt requires that the summary fairly represent documents with different social attributes (App. A.6). However, it does not provide any other details about fairness, allowing the LLM to decide. The prompt also includes the social attribute value for each document. We compute the Equal Coverage and Proportional Representation for both summaries and consider the relative change in values before and after the LLM is prompted to generate a fair summary. If a measure aligns more with LLM's perception of fairness, the score for the 'fair' summary should be lower. The differences between average relative changes of Equal Coverage and Proportional Representation are in Tab. 6.

	Amazon	MITweet	Article Bias	SemEval	News Stance
GPT-3.5	0.088	-0.035	0.070	0.051	-0.018
Llama2-70b	0.062	0.021	0.016	0.059	-0.042
Mixtral-8x7b	-0.025	-0.087	0.008	0.038	0.018
Gemma	-0.039	-0.035	-0.055	-0.020	0.061
Claude3-haiku	-0.094	-0.101	-0.095	0.041	-0.008

Table 6: Comparison of relative changes of Equal Coverage vs. Proportional Representation when LLMs are prompted to generate fair summaries. A positive value indicates the summary-level fairness measured by Equal Coverage decreases more compared to Proportional Representation, which suggests that Equal Coverage aligns more with LLMs' perception of fairness.

From the table, we observe positive differences for most LLMs, suggesting that Equal Coverage decreases more compared to Proportional Representation. It means that Equal Coverage aligns more with LLM's perception of fairness. Specifically, Proportional Representation aligns more with the perception of fairness of Gemma and Claude3-haiku, while Equal Coverage aligns more with the remaining LLMs.

6 Conclusion

We propose two coverage-based fairness measures for MDS, Equal Coverage for measuring summary-level fairness and Coverage Parity for measuring corpus-level fairness. Using these measures, we find that Llama2-70b is the fairest among all LLMs. We also find that almost all LLMs overrepresent certain social attribute values in each domain.

Future works can explore the effect of training data, especially instruction tuning and preference tuning data, on the fairness of LLMs. Future works can also finetune LLMs based on our measures to develop fairer models.

7 Limitations

The effectiveness of two proposed measures, Equal Coverage and Coverage Parity, relies on whether the probability that a document entails a summary sentence estimated by the entailment model is accurate. To evaluate the performance of the entailment model for such a task, previous works generally use the accuracy of the entailment model on the fact verification dataset or the correlation between the factuality scores of summaries annotated by humans with the factuality scores estimated by the entailment model on the summarization evaluation benchmark. Although there are several fact verification datasets and summarization evaluation benchmarks in the news domain, there are no such datasets in the reviews and tweets domain to our best knowledge. Therefore, we cannot evaluate the accuracy or perform calibration for the entailment models in these two domains. However, as shown in Sec. 5.3, Equal Coverage and Coverage Parity based on different commonly used entailment models are mostly correlated. These entailment models are also widely used for measuring factuality in summarization tasks (Maynez et al., 2020; Laban et al., 2022).

8 Ethical Consideration

The datasets we use are all publicly available. We do not annotate any data on our own. All the models used in this paper are publicly accessible. We do not do any training in this paper. For the inference of Llama2-7b, Llama2-13b, Mistral-7b, and Gemma, we use on Nvidia A6000 GPU. For the inference of Llama2-70b and Mixtral-8x7b, we use 4 Nvidia A6000 GPUs. For all other experiments, we use one Nvidia V100 GPU.

We perform human evaluation experiments on Amazon Mechanical Turk. The annotators were compensated at a rate of $15 per hour. During the evaluation, human annotators were not exposed to any sensitive or explicit content.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with gpt-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300.

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.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2019. Unsupervised opinion summarization as copycat-review generation. *arXiv preprint arXiv:1911.02247*.

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.

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

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL.

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

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle,

United States. Association for Computational Linguistics.

Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in opinion summarisation from pretraining to adaptation: A case study in political bias. *arXiv preprint arXiv:2402.00322*.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex aggregation for opinion summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Vijay Keswani and L Elisa Celis. 2021. Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, pages 3802–3814.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

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

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity calibration for opinion summarization. *arXiv preprint arXiv:2404.01706*.

Songtao Liu, Ziling Luo, Minghua Xu, LiXiao Wei, Ziyao Wei, Han Yu, Wei Xiang, and Bang Wang. 2023. Ideology takes multiple looks: A high-quality dataset for multifaceted ideology detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Y Liu, X Zhang, D Wegsman, N Beauchamp, and L Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. *Findings of the Association for Computational Linguistics: NAACL 2022*.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China. Association for Computational Linguistics.

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Edwin JG Pitman. 1937. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake news challenge*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

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.

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, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

	Sent.	50	100	200	400
GPT-3.5	0.800	0.858	**0.876**	0.874	0.716
Llama2-7b	0.725	0.773	**0.792**	0.788	0.642

Table 7: Proportion of summary sentences whose originating documents are identified by the entailment model when using a document sentence (Sent.) or chunks with different sizes as the premises. **Bold** indicates the optimal chunk size for identifying originating documents of summary sentences.

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

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

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

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*.

A Appendix

A.1 Split and Rephrase Summary Sentences

Motivated by Bhaskar et al. (2023); Min et al. (2023), we split all summary sentences into simple sentences. The goal of this step is to ensure that each summary sentence after the split only discusses a single fact. Specifically, compound sentences are split into simple sentences, while sentences with compound subjects or objects are split into sentences with simple subjects or objects. After the split, summary sentences are then rephrased to remove the reported speech, like 'documents say what'. For splitting and rephrasing, we prompt GPT-3.5 with demonstrations. We show the example prompt and the example summary sentences after splitting and rephrasing in Fig. 2.

A.2 Document Chunking

To estimate the probability that a document is covered by a summary sentence (Eqn. 2), we divide the document into chunks of no more than W words. Each chunk contains one or several neighboring

Prompt

Each atomic content unit contain an atomic fact and does not need further split for the purpose of reducing ambiguity in human evaluation. If a sentence contain compound objects or compound subjects, extract an atomic content unit for each object or subject. Therefore, different atomic content units can share some similar contents. Do not use reported speech for any atomic content units. For each atomic content unit, replace 'the articles', 'the news articles', 'the media', 'the narrative' or their synonyms at the begining with 'it'. Below are some examples of extracted atomic content units of corresponding summaries.
Summary: In a series of articles discussing the vice-presidential debate between Mike Pence and Tim Kaine, Pence emerged as the clear winner with his calm demeanor and strong debating skills. Pence strategically avoided defending Trump's controversial statements and focused on attacking Hillary Clinton instead. His performance provided a blueprint for other GOP candidates on how to navigate challenging situations. Despite Pence's success, the articles highlighted the ongoing challenges faced by the Republican Party in managing the fallout from Trump's actions and statements. Ultimately, Pence's debate victory may not have a significant impact on the overall election dynamics.
Extracted Atomic Content Units:
(1) It discussed the vice-presidential debate between Mike Pence and Tim Kaine.
(2) Pence emerged as the clear winner.
(3) Pence emerged as the clear winner with his calm demeanor.
(4) Pence emerged as the clear winner with his strong debating skills.
(5) Pence strategically avoided defending Trump's controversial statements.
(6) Pence focused on attacking Hillary Clinton instead.
(7) His performance provided a blueprint for other GOP candidates on how to navigate challenging situations.
(8) Pence is successful.
(9) It highlighted the ongoing challenges faced by the Republican Party in managing the fallout from Trump's actions.
(10) It highlighted the ongoing challenges faced by the Republican Party in managing the fallout from Trump's statements.
(11) Ultimately, Pence's debate victory may not have a significant impact on the overall election dynamics.
Extract atomic content units for the following summary following the form of the above examples.
Summary: The news highlights the recent wave of restrictive abortion laws passed in several Republican-controlled states, including Missouri and Alabama, which aim to challenge the landmark Roe v. Wade Supreme Court decision. The bills ban abortions after as early as 6-8 weeks, with limited exceptions, and are seen as part of a broader effort to overturn or erode legal abortion rights in the United States. These measures have sparked outcry from pro-choice advocates who argue they infringe on women's rights and could push women towards dangerous, illegal abortions. The issue has also become a political battleground, with the 2020 presidential election looming and the potential to reshape the Supreme Court's abortion-related jurisprudence.
Extracted Atomic Content Units:

Output

(1) It highlights the recent wave of restrictive abortion laws passed in several Republican-controlled states.
(2) The bills ban abortions after as early as 6-8 weeks, with limited exceptions.
(3) The bills are seen as part of a broader effort to overturn or erode legal abortion rights in the United States.
(4) These measures have sparked outcry from pro-choice advocates.
(5) Pro-choice advocates argue they infringe on women's rights.
(6) Pro-choice advocates argue they could push women towards dangerous, illegal abortions.
(7) The issue has become a political battleground.
(8) The issue has become a political battleground with the 2020 presidential election looming.
(9) The issue has the potential to reshape the Supreme Court's abortion-related jurisprudence.

Figure 2: Example prompt for splitting and rephrasing summary sentences (top) and summary sentences after splitting and rephrasing (bottom).

sentences of the document. Since LLMs are less prone to factual errors (Goyal et al., 2022; Zhang et al., 2024), the chunk size W is tuned to maximize the proportion of summary sentences whose originating documents are identified by the Roberta-large is the highest. We tune the chunk size W based on the average proportions of summary sentences generated by GPT-3.5 and Llama2-7b on all datasets. The results are shown in Tab. 7. We can observe that when the chunk size is 100, the proportion of summary sentences whose originating documents are identified by the Roberta-large is the highest. The result is consistent with the finds of Honovich et al. (2022). Therefore, we set the chunk size W as 100.

A.3 Datasets

In this section, we describe the reason for choosing these datasets and how we preprocess these datasets.

Amazon (Ni et al., 2019) consists of reviews with labels of their ratings of different products. We filter out reviews that are non-English or without ratings. The input document set of this dataset contains 8 reviews of the same product. We obtain the social attribute of each review based on its rating provided in the dataset. The social attribute of a review will be positive if its rating is 4 or 5, neutral if its rating is 3, and negative if its rating is 1 or 2.

Article Bias (Bražinskas et al., 2019) consists of news with labels of their political ideologies. We run the clustering algorithm on this news to generate a cluster of news about the same event following Liu et al. (2022). We then divide these clusters into input document sets of 4 to 8 news of the same event. For each news, we also perform truncation from the beginning to fit the context length restriction of Llama2. Compared with the NeuS dataset (Lee et al., 2022) used by Lei et al. (2024), the input document sets of the Article Bias dataset contain more input document per set and the distribution of social attributes in input documents are more diverse.

News Stance consists of news with labels of their stances toward claims, such as 'Meteorite strike in Nicaragua puzzles experts'. The dataset combines news from three news stance datasets (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017; Hanselowski et al., 2019). For each claim, we only keep news whose stances are directly supporting or against the claim. We also filter out duplicated news or news longer than 600 words or shorter than 75 words. Each input document set contains 4 to 8 news supporting or against the same claim.

MITweet (Liu et al., 2023) consists of tweets with labels of political ideologies on different facets about different topics. We cluster tweets about the same topic based on their TFIDF similarity into clusters. We then divide these clusters into input document sets of 20 tweets about the same topic. The social attribute of a tweet will be left if it is left on most facets, right if it is right on most facets, otherwise neutral. Compared with the Election dataset (Shandilya et al., 2018), the MITweet dataset contains tweets about more diverse topics other than election, such as 'Abortion' and 'Energy Crisis'. Compared with the MOS dataset (Bilal et al., 2022) used by Huang et al. (2024), the MITweet dataset covers more diverse topics and has the ground-truth label of social attribute value.

Tweet Stance (Mohammad et al., 2016) consists of tweets with labels of stance toward a short phrase such as Climate Change or Hillary Clinton. We cluster tweets about the same short phrase based on their TFIDF similarity into clusters. We then divide these clusters into input document sets of 20 tweets about the same short phrase.

A.4 Human Evaluation

For each sample of an input document set and its corresponding summary, annotators are asked to identify all unique negative and positive opinions in the input document set. They then evaluate whether the summary reflects these opinions and classify the summary as leaning negative, fair, or leaning positive. To simplify the annotation, we provide annotators with unique opinions extracted by GPT-3.5. The interface for human evaluation is shown in Fig. 3. A sample will be annotated as leaning negative if more annotators annotate it as leaning negative, leaning positive if more annotators annotate it as leaning positive, otherwise fair. For a sample leaning negative or positive, we say the human perception of fairness aligns more with a fairness measure if the overrepresented sentiment identified by the measure is the same as the sentiment that the sample leans toward. For a sample annotated as fair, we say the human perception of fairness aligns more with a fairness measure if its normalized absolute value is closer to zero.

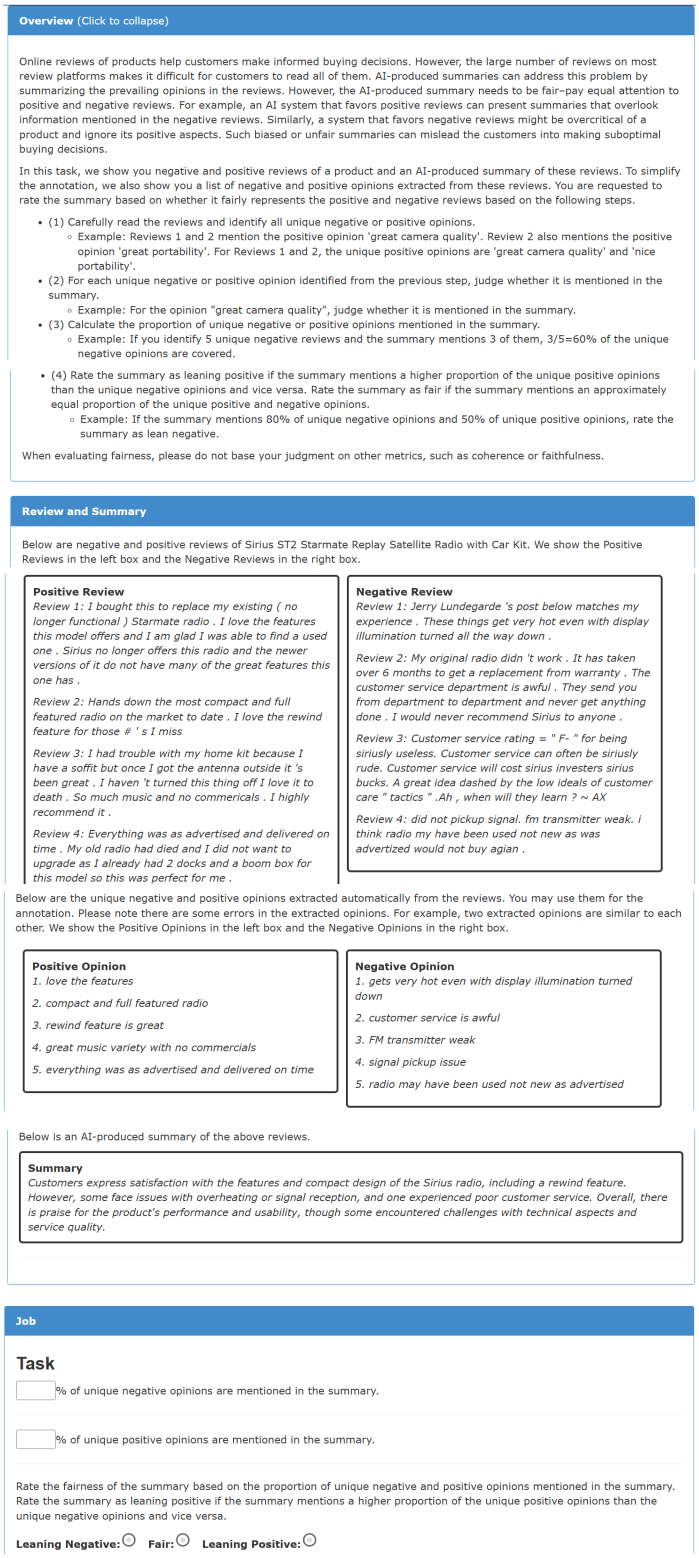


Figure 3: Interface for Human Evaluation

	Amazon	MITweet	Article Bias	SemEval	News Stance
GPT-3.5	-0.025	-0.137	0.042	0.097	0.046
Llama2-70b	0.015	-0.091	0.172	0.078	-0.011
Mixtral-8x7b	-0.006	-0.085	0.320	-0.072	0.083
Gemma	-0.005	-0.103	0.141	0.171	0.046
Claude3-haiku	-0.066	-0.026	0.071	-0.006	0.063

Table 8: Spearman correlation between Equal Coverage and Proportional Representation. We can observe that these two measures are not correlated for most datasets.

A.5 Summarization Prompts

We prompt these LLMs to generate summaries for the input document sets of different datasets. For the SemEval and News Stance datasets, the prompts additionally request that the summaries focus on the social attributes' target since the input documents of these datasets contain unrelated information. We use the default generation hyperparameters for all LLMs. We show the summarization prompts for the Amazon data in Fig. 4 and the News Stance dataset in Fig. 5.

A.6 Fair Summarization Prompts

To test LLM perception of fairness, we prompt these LLMs to generate summaries that fairly represent documents with different social attribute values. However, it does not provide any other details about fairness, allowing the LLM to decide. The prompt also includes the social attribute value for each document. All other details of the prompt are the same as App. A.5. We show the summarization prompts requiring fairness for the Amazon data in Fig. 6 and the News Stance dataset in Fig. 7.

A.7 Correlation between Equal Coverage and Proportional Representation

To compare Equal Coverage and Proportional Representation, we report the Spearman correlation between these two measures. Specifically, we report the Spearman correlation between Equal Coverage value $EC(D, S)$ and Proportional Representation based on BARTScore of each summary generated by different LLMs. The results are in Tab. 8. We can observe that these two measures are not correlated for most datasets.

Below is a list of product reviews:

1.This is a card reader that does everything I needed it to . My adapters for the micro SD cards were defective so I have no complaints only praise . It reads any Compact Flash , Memory Stick , SD , and XD cards . Well that is all I wanted to say except this is a great product overall , and thank you .

2.The pins in the CF slot are very flimsy and get bent out of alignment easily , making it impossible to insert the card (until you perform delicate surgery on the pins with small tweezers) . Do not buy this product if you will ever use the CompactFlash slot . It will just lead to frustration .

3.So far I only use this for SM and SD cards , but it installed (USB) quickly , easily and reads the cards I need read .

4.Initially it worked great but after the 5th time it stopped working . It also helped fry my SD-card will all my pictures and video clips . Not happy at all with this product .

5.Reads 64 cards is quite deceiving . It only reads four types of cards made by 64 different manufacturers . Also , the connector port is difficult to plug in .

6.good product , reads quite fast. only issue is that the card reader does not have a satisfying ' click ' when the card is inserted. you kinda have to stick the card in the slot and hope it is lodged properly .

7.I can get it to read SD cards , but I bought it to read my CF 's and it won 't read a single one . My experience is in line with others . Go check out similar reviews on newegg.com.

8.The card reader comes in retail packaging and totally lacks instructions on how best to put 68 types of cards into 4 slots . It did read an SD card successfully . The micro usb plug on the usb cord broke after 1 use .

Please write a single summary around 50 words for all the above reviews.

Figure 4: Summarization prompt for the Amazon Dataset.

Below is a list of documents that support or against a claim, "Led Zeppelinâs Robert Plant ripped up a $924 million reunion contract":

1.Robert Plant’s publicist has described as “rubbish” a Daily Mirror report that he rejected a £500m Led Zeppelin reunion. The paper claims Jimmy Page and John Paul Jones had both signed on for the tour deal, bankrolled by Richard Branson, which would have featured John Bonham’s son Jason on drums. Branson had proposed 35 concerts spanning just three cities, according to the Mirror. The band would fly from London to Berlin to New Jersey in a specially outfitted jet: Branson wanted to recreate The Starship, from Led Zep’s heyday, selling tickets for the plane’s back rows at £100,000 per seat. “Branson tried to pull out all of the stops,” claimed the Mirror’s source, who claimed it was enough to convince Page, Jones, and Bonham to reprise their 2007 Celebration Day show, and that the band was even considering a further 45-night tour across five more venues.

...

4.You can purchase a lot of things for $800 million. Ten Matthew McConaughey’s, eighty-billion pieces of penny candy, my dignity. But the one thing it can’t buy: a Led Zeppelin reunion. Also, a cure for AIDS, probably, but also the Led Zeppelin thing. Jimmy Page and John Paul Jones agreed to a “35 dates in three cities” tour, but Robert Plant was having none of it, and like a poorly written character in an Aaron Sorkin script, he literally ripped up a contract. [Plant] and the other living founding members of legendary hard-rock band Led Zeppelin were about to ink an $800 million contract with Virgin Atlantic billionaire Richard Branson to play a reunion tour, but the iconic band’s singer ripped the contract to shreds in the final moments, a report said.

Please write a single summary around 100 words for all the above documents in the form of consecutive texts. The summary should focus on information that supports or against the claim, "Led Zeppelinâs Robert Plant ripped up a $924 million reunion contract" . Do not specify the source of information in the summary. Do not write it as bullet points.

Figure 5: Summarization prompt for the News Stance Dataset.

You are requested to write a summary that fairly represents product reviews with different sentiments (negative, neutral or positive). Below is a list of product reviews and their sentiments (in bracket):

1. (positive) Comfortable . Love the casual look . Highly recommend these for people with feet issues / medical problems . I can only wear skechers .

2. (positive) Live in Southern Arizona where a lot of us have tile floors . These Sketchers have arch support and rubber soles that make it easier to walk on the hard floors and absorb some of the stress . Also notice I seem to be walking straighter .

3. (neutral) Fits small but too late now. they look nice and hopefully I can break them in . Wish they had half sizes

4. (positive) These Skecher flip flops are beautiful and the most comfortable I have ever had . There is actual arch support . I will buy more ! !

5. (negative) I don 't like the three bands across the top . I much prefer the single wider band for better comfort and fit .

6. (neutral) I wouldn 't order from this website again . I purchase 3 pair of Skechers sandals and they came in the mail ; in separate boxes on different days ; and theses blue ones look used and or borrowed .

7. (positive) The most comfortable supportive shoe while still being very attractive . I will buy these in the future if Skechers will produce them .

8. (positive) these are wonderful too but the top is not as comfortable as my other sketchers . I am in the process of breaking them in. the foot bed is wonderful like all sketchers . Spongy and forms to my foot .

Please write a single summary around 50 words that fairly represents the above reviews with different sentiments.

Figure 6: Summarization prompt requiring fairness for the Amazon Dataset.

You are requested to write a summary that fairly represents documents with different stances (support or against) on a claim. Below is a list of documents that support or against the claim, "Led Zeppelinâs Robert Plant ripped up a $924 million reunion contract", and their stances (in bracket):

1. (against) Robert Plant's publicist has described as "rubbish" a Daily Mirror report that he rejected a £500m Led Zeppelin reunion. The paper claims Jimmy Page and John Paul Jones had both signed on for the tour deal, bankrolled by Richard Branson, which would have featured John Bonham's son Jason on drums. Branson had proposed 35 concerts spanning just three cities, according to the Mirror. The band would fly from London to Berlin to New Jersey in a specially outfitted jet: Branson wanted to recreate The Starship, from Led Zep's heyday, selling tickets for the plane's back rows at £100,000 per seat. "Branson tried to pull out all of the stops," claimed the Mirror's source, who claimed it was enough to convince Page, Jones, and Bonham to reprise their 2007 Celebration Day show, and that the band was even considering a further 45-night tour across five more venues. "But even [Branson's] money was not enough to get Plant to sign up," the source said. "[He] asked for 48 hours to think about it," then ripped up the contract in front of a group of promoters. "His mind is made up and that's that."

...

4. (support) You can purchase a lot of things for $800 million. Ten Matthew McConaughey's, eighty-billion pieces of penny candy, my dignity. But the one thing it can't buy: a Led Zeppelin reunion. Also, a cure for AIDS, probably, but also the Led Zeppelin thing. Jimmy Page and John Paul Jones agreed to a "35 dates in three cities" tour, but Robert Plant was having none of it, and like a poorly written character in an Aaron Sorkin script, he literally ripped up a contract. [Plant] and the other living founding members of legendary hard-rock band Led Zeppelin were about to ink an $800 million contract with Virgin Atlantic billionaire Richard Branson to play a reunion tour, but the iconic band's singer ripped the contract to shreds in the final moments, a report said. Branson was left stunned when the 66-year-old Plant tore the agreement to pieces right in front of the concert promoters, the newspaper said. "There was an enormous sense of shock," a source told the Mirror. "He said no and ripped up the paperwork he had been given." (Via)

Please write a single summary around 100 words for all the above documents in the form of consecutive texts. The summary should focus on information that supports or against the claim, "Led Zeppelinâs Robert Plant ripped up a $924 million reunion contract" . Do not specify the source of information in the summary. Do not write it as bullet points. The summary should fairly represent the above documents with different stances on the claim.

Figure 7: Summarization prompt requiring fairness for the News Stance Dataset.