Fair Abstractive Summarization of Diverse Perspectives

Yusen Zhang[♣] **Nan Zhang**[♣] **Yixin Liu**[†]
Alexander Fabbri[♢] **Junru Liu**[♠] **Ryo Kamoi**[♣] **Xiaoxin Lu**[♣]
Caiming Xiong[♢] **Jieyu Zhao**[‡] **Dragomir Radev**[†] **Kathleen McKeown**[♡] **Rui Zhang**[♣]
[♣]Penn State University [†]Yale University [♢]Salesforce Research
[♠]Texas A&M University [‡]University of Southern California [♡]Columbia University
{yfz5488, rmz5227}@psu.edu

Abstract

People from different social and demographic groups express diverse perspectives and conflicting opinions on a broad set of topics such as product reviews, healthcare, law, and politics. A fair summary should provide a comprehensive coverage of diverse perspectives without underrepresenting certain groups. However, current work in summarization metrics and Large Language Models (LLMs) evaluation has not explored fair abstractive summarization. In this paper, we systematically investigate fair abstractive summarization for user-generated data. We first formally define fairness in abstractive summarization as not underrepresenting perspectives of any groups of people, and we propose four reference-free automatic metrics by measuring the differences between target and source perspectives. We evaluate nine LLMs, including three GPT models, four LLaMA models, PaLM 2, and Claude, on six datasets collected from social media, online reviews, and recorded transcripts. Experiments show that both the model-generated and the human-written reference summaries suffer from low fairness. We conduct a comprehensive analysis of the common factors influencing fairness and propose three simple but effective methods to alleviate unfair summarization. Our dataset and code are available at https://github.com/psunlpgroup/FairSumm.

1 Introduction

Different social and demographic groups of people hold diverse and potentially even conflicting perspectives and opinions, which are expressed in user-generated text data in various domains and topics such as product reviews, social debates, healthcare, law, and politics (Shandilya et al., 2018; Bražinskas et al., 2020a; Zhang et al., 2016; Huang et al., 2023b,a; Kovač et al., 2023; Hayati et al., 2023). When a summarization system is faced with diverse perspectives that can be equally correct and fundamental, a fair summary shall provide an accurate and comprehensive view of diverse perspectives from these groups. For example, Figure 1 shows the reviews about an Amazon product. The summary generated by GPT-3.5 (Ouyang et al., 2022) unfairly represents sentiments because it summarizes more content from positive reviews while ignoring some critical points from negative reviews.

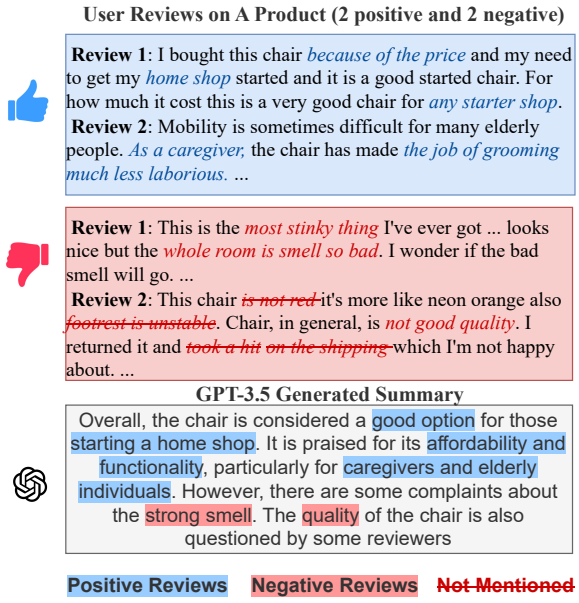

Figure 1: An example from PERSPECTIVESUMM. The blue/red box displays the input consisting of positive/negative reviews. The grey box shows the summary generated by GPT-3.5 (text-davinci-003). The generated summary is unfair because the negative reviews are underrepresented compared with the positive reviews.

However, existing summarization metrics cannot adequately measure fairness. Metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) fail to measure fairness for two reasons. First, they are not inherently designed for quantifying fairness, but rather measuring similarity between system outputs and references based

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3404–3426

on n-gram overlapping or embeddings. Second, reference summaries can also be biased. Moreover, current work on summarization (Shandilya et al., 2018) and LLM evaluation (Chang et al., 2023b) has not explored the fairness of abstractive summarization, which is more prevalent than extractive summarization in the era of LLMs (OpenAI, 2023).

In this work, we present the first study that systematically *defines*, *quantifies*, and *benchmarks* the fairness of abstractive summarization using large language models in a wide range of domains and topics. First, we define our task of fair abstractive summarization as generating summaries that are fair with respect to a social attribute that can take different values, such as the sentiment attribute with positive/negative values. We curate a benchmark PERSPECTIVESUMM (Table 1) by unifying and cleaning six existing datasets in domains where fairness is a critical issue including healthcare, politics, restaurant, product, law, and policy, which covers social attributes of gender, party, sentiment, rating, and speaker. The fairness of the summary on these attributes greatly influences the information and even the opinions of the readers. Next, we quantify fairness in abstractive summarization as a distribution alignment between the generated summary and the source text. The distribution is calculated by ratios of semantic units of different social attribute values. According to the definition, we propose several metrics to evaluate fairness, including Binary Unfair Rate (BUR) which determines unfair summaries by checking if any social attribute value is underrepresented, Unfair Error Rate (UER) which measures the distance between the generated summary and the source text in terms of their social attribute value distributions. Third, we benchmark fairness in abstractive summarization by using PERSPECTIVESUMM to conduct a comprehensive evaluation of nine main-stream LLMs, including GPT (text-davinci-003 (Brown et al., 2020), gpt-turbo-3.5 (Ouyang et al., 2022), gpt-4 (OpenAI, 2023)), LLaMA (Alpaca (Taori et al., 2023), llama2-chat-7/13/70B (Touvron et al., 2023b)), PaLM 2 (Anil et al., 2023) (text-bison@001), and Claude (Anthropic, 2023) (claude-instant-1).

Results of our metrics and human evaluation show that neither human-written reference summaries nor LLM-generated summaries could maintain fairness based on our definition. Even GPT-4 suffers from fairness problems as the BUR score of GPT-4 on Amazon reviews reaches 74.27%. We also provide three simple but effective ways to improve the summary fairness including changing decoding temperature, changing summary length, and appending the definition of fairness to the instruction prompt. The result shows that BUR decreases from 51.11% to 47.48% on gpt-turbo-3.5 by providing instructions of fairness definition.

Our contributions are as follows: (1) We propose new metrics for quantifying and measuring the fairness of abstractive summarization over diverse perspectives. (2) We collect and unify a comprehensive benchmark for analyzing fairness of abstractive summarization in various domains and topics. (3) We conduct comprehensive experiments to investigate the fairness of reference summaries and summaries generated by popular LLMs, finding that both reference/generated summaries often fail to achieve fairness. (4) We analyze the factors that influence fairness to propose three simple but effective methods to improve fairness.

2 Summarization of Diverse Perspectives with Social Attributes

In this section, we first introduce social attributes, then formulate the task of summarization from diverse perspectives with social attributes.

Social Attributes. A fair abstractive summary should include comprehensive perspectives without underrepresenting certain groups of people. We use social attributes to indicate the properties that form groups of people. As shown in Table 1, social attributes can be gender, party, and sentiment.

Definition 2.1 Summarization with Social Attribute. A summarization dataset $\mathcal{D}$ consists of pairs of source text and target summary $(\boldsymbol{x}, \boldsymbol{y})$. Here, $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]$ is the source text consisting of n token x_i and $\boldsymbol{y} = [y_1, y_2, \cdots, y_m]$ is the target summary consisting of m tokens y_i. We consider a social *attribute* a having r different values $\mathcal{V} = \{v_1, v_2, \cdots, v_r\}$, $r > 1$. Furthermore, we use $v(\cdot)$ to denote the value of a token. We assume each source unit x_i belongs to one value $v(x_i) \subset \mathcal{V}$, $|v(x_i)| = 1$, while each token in target is associated with one or multiple values $v(y_i) \subseteq \mathcal{V}$, $|v(y_i)| \geq 1$. For instance, when considering gender as an attribute a in Twitter review summarization, the $r = 2$ possible values are $\mathcal{V} = \{\texttt{male}, \texttt{female}\}$. For each tweet token x_i in the source, if it is written by a male, $v(x_i) = \texttt{male}$,

otherwise $v(x_i) = \texttt{female}$[1]. For each token y_i in the summary of tweets, it could come from `male`, `female`, or both. For instance, if both groups claim a product is "good". Then "good" can be both from `male`, and `female`.

3 Definition of Fairness in Summarization

In this section, we provide definitions for fairness. Figure 2 shows the logic of computation of fairness in abstractive summarization and corresponding evaluation metrics. The first stage is to produce the value distribution. Given a sample pair of source and target text, we first split the source according to the annotated values of social attributes. Then, we apply N-gram/BERT/BART scores to compare target tokens with source tokens to obtain the value of each token in the summary. Next, we count the number of tokens for each value in the source and target and obtain the value distribution of them.

3.1 Defining Value Distribution

Computation of source and target value distribution (source/target distribution for short) is the first step of fairness computation. The formal definition is listed as follows.

Definition 3.1 Value Distribution. We define the value distribution $\boldsymbol{p}$ in three scenarios: (1) the value distribution in source: $\boldsymbol{p_x} = [\boldsymbol{p_x}(v_1), \boldsymbol{p_x}(v_2), \cdots, \boldsymbol{p_x}(v_r)]$, where $\boldsymbol{p_x}(v_k)$ is the proportion of source tokens associated with value v_k, $0 \leq \boldsymbol{p_x}(v_k) \leq 1$, $\sum_{k=0}^{r} \boldsymbol{p_x}(v_k) = 1$. (2) Similarly, the value distribution in target: $\boldsymbol{p_y} = [\boldsymbol{p_y}(v_1), \boldsymbol{p_y}(v_2), \cdots, \boldsymbol{p_y}(v_r)]$. (3) a gold value distribution $\boldsymbol{p_g} = [\boldsymbol{p_g}(v_1), \boldsymbol{p_g}(v_2), \cdots, \boldsymbol{p_g}(v_r)]$ which means the distribution that aligned with user's requirement (Section 3.3).

3.2 Calculating Value Distribution

To compute $\boldsymbol{p_x}$, since the meta-data of the dataset shows the values of each token in the source, it can be acquired by counting the number of tokens of each value. However, the calculation of $\boldsymbol{p_y}$ cannot be obtained easily due to the abstract nature of summaries. As abstractive summarization does not directly copy from the source text, it contains words that are not in the source but still semantically come from the source. This makes it difficult to attribute a sentence in the summary to a certain part of the source. To tackle this, we propose two algorithms for calculating the target distribution $\boldsymbol{p_y}$:

N-gram Matching. For a given k, we scan each k-gram in target text, if this k-gram is also shown in the source text, we assign this k-gram the value of the source text it appears. It is worth noting that the k-gram in the target can be assigned multiple values according to the definition. These metrics can automatically compute $\boldsymbol{p_y}$ given $\boldsymbol{p_x}$, with no additional annotation on $\boldsymbol{p_y}$. We use $k = 1$ for all experiments. We use N-gramScore to represent this matching approach.

Neural Matching. We use BARTScore (Yuan et al., 2021), and BERTScore (Zhang et al., 2020) to measure the distance between the target and source. These metrics capture the similarity in semantics, overcoming the new word matching issues in the n-gram algorithm (Banerjee and Lavie, 2005). Specifically, for a given source $\boldsymbol{x}$ and target $\boldsymbol{y}$, we first split $\boldsymbol{x}$ by value. For each $v_i \in \mathcal{V}$, $\boldsymbol{x_i} = \{x \in \boldsymbol{x} | v(x) = v_i\}$. Let $r = |\mathcal{V}|$. Then, we compute the correlation between $\boldsymbol{x_i}$ and $\boldsymbol{y}$ to obtain $\boldsymbol{p_y}$: $\boldsymbol{p_y} = \text{Softmax}(\text{Score}(\boldsymbol{x_1}, \boldsymbol{y}), \cdots, \text{Score}(\boldsymbol{x_r}, \boldsymbol{y}))$, where Score is BARTScore and BERTScore, and the temperature of Softmax is a hyperparameter. We pick 0.1 for experiments because it better aligns with human evaluation (Appendix 6.4).

The design of the metrics is different in terms of the following two aspects. First, they have different granularity. N-gramScore is based on the token which controls the fairness of the token level. BERTScore computes the similarity of the sentences which is in sentence level. BARTScore computes via the entailment of the entire summary. This is on the summary level. Second, these scores have different advantages. N-gramScore can be applied to diverse length source text while the other two scores have length limitations. BERT score can capture semantic similarity while BART score captures entailment. These two models rely on the accuracy of the backbone models as well.

3.3 Defining Summarization Fairness

After defining the source/target distribution, the goal of fair summarization is equivalent to producing a $\boldsymbol{p_y}$ not underrepresenting any value in $\boldsymbol{p_x}$.

Definition 3.2 Summarization Fairness. We define an unfair summary as exhibiting an underrepresentation of user groups in summaries. Given value distributions, we can define several types

[1]To simplify the task setting, we consider two genders in our experiments, while our approach can generalize.

Dataset	Domain	Source Form	Collect From	Attributes	Values	# Samples	Max and Avg Length
Claritin	Healthcare	Social media	twitter.com	Gender	2	1350	1087/572.2
US Election	Politics	Social media	twitter.com	Party	3	1350	1247/677.7
Yelp	Restaurant	Review	yelp.com	Sentiment	3	1500	576/402.1
Amazon	Product	Review	amazon.com	Rating	5	1500	531/346.5
Supreme Court	Law	Dialogue	supremecourt.gov	Speakers	3-10	665	2721/1910.9
Intelligence Squared (IQ2)	Public Policy	Debate	opendebate.org	Speakers	2-10	1421	3275/1650.7

Table 1: Dataset characteristics of PERSPECTIVESUMM for fair abstractive summarization. Examples in Table 8.

of fairness: (1) **Ratio Fairness**: the target value distribution $\boldsymbol{p_y}$ shall follow the same value distribution as source $\boldsymbol{p_x}$. In this case, $\boldsymbol{p_g} = \boldsymbol{p_x}$. (2) **Equal Fairness**: the target value distribution $\boldsymbol{p_y}$ shall follow the uniform value distribution $\boldsymbol{p_g} = [1/r, 1/r, \cdots, 1/r]$, regardless of source. This distribution keeps the balance of each value. (3) Furthermore, users can define any $\boldsymbol{p_g}$ to indicate their ideal distribution, not restricted to Ratio Fairness or Equal Fairness. Without loss of generality, we will discuss ratio fairness in our metric and experiment sections, because summarization aims to compress but still follow the original distribution in the source text (Shandilya et al., 2018).

4 PERSPECTIVESUMM

To conduct a comprehensive analysis of fairness, we select and then process existing datasets to form our fair summarization benchmark PERSPECTIVESUMM. We follow two main principles to choose six datasets from a large variety of existing datasets. First is the *quality*, the source texts need to be human-written and marked with clear, and precise social attributes. These attributes are existing metadata in the datasets, except for the sentiment attribute, which is derived from classifier predictions. Second is *diversity*, the collected datasets need to cover various domains and perspectives. Table 1 shows the statistics of PERSPECTIVESUMM. It covers six different domains with six attributes. Overview of these six datasets are listed as follows (details in Appendix A):

Claritin (Shandilya et al., 2018) contains tweets about the effects of the drug Claritin where a = gender and $\mathcal{V} = \{\text{male}, \text{female}\}$. **US Election** (Shandilya et al., 2018) contains tweets posted during the 2016 US Presidential election where a = politics and $\mathcal{V} = \{\text{pro-rep}, \text{pro-dem}, \text{neu}\}$, meaning pro-republic, pro-democracy, or neutral. For Claritin and US-Election, we randomly sample tweets of different values with a certain ratio and then combine them as one input source.

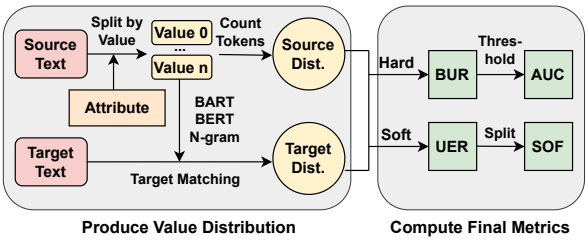

Figure 2: Overview of our proposed metrics. Dist. means value distribution.

Amazon and **Yelp** are two datasets from the FewSum dataset (Bražinskas et al., 2020a) containing product and business reviews. Each sample in these two datasets consists of eight user-written reviews. In Amazon dataset, a = ratings, and $\mathcal{V} = \{1, 2, 3, 4, 5\}$ shows the ratings of each reviewer. In Yelp dataset, a = sentiment, and $\mathcal{V} = \{\text{pos}, \text{neu}, \text{neg}\}$ displays the sentiment of reviews. **SupremeCourt** (Danescu-Niculescu-Mizil et al., 2012) contains a collection of conversations from the U.S. Supreme Court Oral Arguments, where a = speakers, $\mathcal{V}$ are the names of all participants. **Intelligence Squared Debate dataset (IQ2)** (Zhang et al., 2016) collects public debates that follow the Oxford style and are recorded live, where a = speakers, $\mathcal{V}$ contains the names of all participants. For SupremeCourt and IQ2, we truncate each transcript of the debate into shorter segments of k tokens. We include the whole text of each sample for the experiments of claude-100k.

5 Evaluating Fairness in Summarization

As shown in Figure 2, we define four types of metrics to evaluate whether the generated summaries are fair: Binary Unfair Rate (BUR), Unfair Error Rate (UER), Area Under Curve (AUC), and Second-Order Fairness (SOF).

Definition 5.1 Binary Unfair Rate (BUR). We define Binary Unfair Rate (BUR), a binary function that outputs 1 if the sample is unfair; and 0 otherwise. To be specific, we define the summary

as fair if and only if $\boldsymbol{p_y}(v_k) \geq \tau \cdot \boldsymbol{p_x}(v_k)$ holds for each value $k = 1, 2, \cdots, r$, where $0 \leq \tau \leq 1$ is a tolerance hyperparameter that is chosen based on application scenarios (Shandilya et al., 2018). Otherwise, if $\boldsymbol{p_y}(v_k) < \tau \cdot \boldsymbol{p_x}(v_k)$, we say the value v_k is underrepresented. Thus

$$f_{\text{BUR}}(\boldsymbol{x}, \boldsymbol{y}; \tau) = \mathbb{1}\left(\bigvee_{i=1}^{r} \boldsymbol{p_y}(v_k) < \tau \cdot \boldsymbol{p_x}(v_k)\right) \quad (1)$$

This metric is either zero to one and it can be averaged over the dataset, demonstrating the proportion of summaries that are *unfair* in the dataset. However, this hard metric may not be able to describe which one is more unfair/fair if BUR of two samples is the same. Thus, we propose Unfair Error Rate as a soft metric to compute the sum of the distance to source distribution of each value.

Definition 5.2 Unfair Error Rate (UER). This metric defines a function $f_{\text{UER}}(\boldsymbol{x}, \boldsymbol{y})$ measuring the distance between $\boldsymbol{p_x}$ and $\boldsymbol{p_y}$. UER serves as the supplementary metric for BUR as it shows more fined-grained details compared with BUR.

$$f_{\text{UER}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{r}\sum_{k=1}^{r} \max(0, \boldsymbol{p_x}(v_k) - \boldsymbol{p_y}(v_k)) \quad (2)$$

UER shows the average percentage of values that are underrepresented. Furthermore, since people have different tolerance and requirements for fairness, we propose Area Under Curve to compute the expected result of BUR when the tolerance is randomly picked from zero to one.

Definition 5.3 Area Under Curve (AUC). As mentioned, τ is a hyperparameter chosen according to the scenarios. However, people may have different assignments of τ according to their own experiences. To incorporate different tolerance hyperparameters, we propose AUC of fairness threshold (AUC for short) for BUR metric $f_{\text{AUC}}(\boldsymbol{x}, \boldsymbol{y})$.

$$f_{\text{AUC}}(\boldsymbol{x}, \boldsymbol{y}) = \int_0^1 f_{\text{BUR}}(\boldsymbol{x}, \boldsymbol{y}; \tau)\, d\tau \quad (3)$$

It is a number between zero and one showing the expected BUR before assigning a fairness threshold τ to it because we assume that the users evenly pick the threshold between zero and one. It is worth noting that, practically, we sample ten points evenly to approximate the integration.

These three metrics cannot show which value is more unfair in one sample. Thus, we further propose Second-Order Fairness to compute the variance of UER among all values.

Definition 5.4 Second-Order Fairness (SOF). Second-Order Fairness (SOF) f_{SOF} measures the unfairness difference among all values, defined as:

$$S_{\text{UER}} = \{\max(0, \boldsymbol{p_x}(v_k) - \boldsymbol{p_y}(v_k)) \mid k \in [r]\}$$
$$f_{\text{SOF}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{r}\sum_{s \in S_{\text{UER}}} \left| s - \frac{\sum_{s \in S_{\text{UER}}} s}{r} \right| \quad (4)$$

Inspired by MacQueen et al. (1967), SOF measures the coherence of UER set S_{UER} via computing the average distance to the center $\sum_{s \in S_{\text{UER}}} s/r$. If all values' UER are close to their averaged center, the possibilities of being unfair are similar for values. In this case, each value has a similar possibility to be unfair, so-called second-order fairness.

Tolerance Hyperparameter. Regarding τ, the threshold that decides whether a value is underrepresented is based on the application scenarios (Celis et al., 2018). When strict fairness is required, such as the political opinions of two parties, this threshold should be set to a high value, for instance, $\tau = 0.8$ or higher (Shandilya et al., 2018). However, in some other text summarization tasks in our daily lives, we do not require each value to have strict constraints, such as product review summarization. In the extreme case, we only want each value to appear in the summary rather than constraining the time of appearance, so we can set $\tau = 0$. In this paper, we set $\tau = 0.8$ for all experiments (Biddle, 2017), if not specified.

Metric Quality Validation. We validate our metrics because they correlate with human evaluation (Appendix 6.4) and achieve lower/upper bound values on extreme synthetic examples (Appendix C).

6 Experiment Results and Analysis

Our results and analysis aim to answer the following research questions:

- RQ 1: How fair are the summaries generated by LLMs based on our metrics (Section 6.1)?
- RQ 2: How fair are the existing human-written reference summaries according to our automatic metrics (Section 6.2)?
- RQ 3: How do humans perceive the fairness of summaries generated by LLMs (Section 6.3) and quality of proposed metrics (Section 6.4)?
- RQ 4: How can we dissect the fairness of summaries generated by LLMs using SOF (Section 6.5) and AUC (Section 6.6)?

		Claritin		US Election		Yelp		Amazon		SupremeCourt		IQ2	
	Size	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓
Alpaca	7B	65.38	10.35	87.33	8.40	44.04	5.11	80.73	7.16	97.99	8.79	91.06	8.99
llama-2-chat	7B	62.99	9.09	**78.64**	6.69	41.76	3.59	76.00	**5.09**	97.40	4.92	**84.41**	**6.02**
llama-2-chat	13B	63.06	9.35	79.28	6.79	**36.20**	3.53	**73.51**	5.11	**95.49**	4.58	86.04	6.66
llama-2-chat	70B	**61.53**	9.92	79.78	**6.41**	37.04	**2.81**	74.69	5.63	96.94	4.74	85.71	6.64
text-bison@001	N/A	67.73	10.49	86.53	8.71	45.65	5.37	84.42	8.43	97.74	5.65	89.26	8.21
text-davinci-003	175B	62.94	**9.08**	82.74	7.09	43.09	4.35	79.89	6.60	97.39	5.25	87.17	7.20
gpt-turbo-3.5	175B	64.30	9.18	81.38	6.48	38.64	4.00	75.89	5.82	96.59	4.64	84.52	6.07
gpt-4	N/A	66.37	9.94	79.93	6.99	39.42	3.72	74.78	5.49	96.79	**4.57**	87.71	6.94

Table 2: Main results with our proposed metrics. We report the mean of N-gramScore, BERTScore, and BARTScore. Full results are in Table 10. BUR and UER are better with a lower score ↓. **Bold** is the best performance, and underline is the second best.

	N-gramScore	
	BUR↓	UER↓
claude-instant-1 (N/A)		
SupremeCourt	99.07	1.77
IQ2	100.00	2.37

Table 3: Long context results. Only N-gramScore is used because the input length exceeds the limit of BERTScore and BARTScore.

	N-gramScore		BERTScore		BARTScore	
	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓
reference summary						
Amazon	95.00	18.50	95.00	18.50	91.67	9.44
Yelp	68.00	26.43	53.04	7.90	56.00	10.11
gpt-turbo-3.5						
Amazon	68.33	3.97	76.67	5.99	93.33	9.51
Yelp	21.00	3.12	36.00	4.76	64.00	6.27

Table 4: Comparison between human-written reference summary and summary generated by gpt-turbo-3.5.

We test nine LLMs as listed in Appendix F. For analysis, we use gpt-turbo-3.5 on the Claritin dataset with N-gramScore as the target matching method, if not specified.

6.1 Overall Results

Table 2 shows the results of all models using our metrics. In general, *our results indicate that many summaries generated by LLMs are not fair*. Most models do not perform well according to our metrics. For example, on Amazon datasets, gpt-4 can achieve a BUR score of 75%. This means around 3 out of 4 summaries of reviews have some bias towards one side or the other. Overall, LLaMAs enjoy better fairness than other models. We found gpt-turbo-3.5 and gpt-4 are in general better than their older version text-davince-003, and 13B/70B llama2-chat are better than their smaller 7B version and Alpaca. However, we don't find strong evidence that gpt-4 is better than gpt-turbo-3.5. Besides, we observe that the performance of gpt-4 and gpt-turbo-3.5 significantly vary on different male ratios and source lengths (Appendix I).

Long Context Results. Different from other models, claude-instant-1 can consume 100k tokens as inputs. Thus, we prepare a full dataset on IQ2 and SupremeCourt without segmentation. As shown in Table 3, compared with short input, it is more difficult to maintain a fair summary. Thus, *Claude also suffers from fairness issues on long input datasets*.

6.2 Fairness of Reference Summary

We also explore the fairness of the human-written reference summaries. We analyze the fairness of 100 reference summaries compared with summaries generated by gpt-turbo-3.5 on Amazon and Yelp datasets. As shown in Table 4, the reference summary also surfers from high BUR and UER scores, indicating that *existing human-written reference summaries cannot maintain the fairness either, even worse than those generated by gpt-turbo-3.5*. This echoes previous findings that LLMs can generate summaries with higher quality than human-written ones (Goyal et al., 2022a).

6.3 Human Evaluation

We present two sets of human evaluation scores, corresponding to the two stages of the annotation process (more design details in Appendix E). In the first stage, we first use a tool provided by Liu et al. (2022) to decompose each sentence into Atomic Content Units (ACUs). Then, we ask the annotators to verify the ACUs and then count the proportions of each value in the summary. We collect the results and compute BUR and UER scores. In a

	Claritin			Yelp			IQ2		
	BUR↓	UER↓	Rating↓	BUR↓	UER↓	Rating↓	BUR↓	UER↓	Rating↓
Alpaca	76.47	13.10	85.00	45.00	4.67	88.24	88.88	11.83	100.00
text-davinci-003	**58.75**	**8.85**	51.76	**15.00**	**2.58**	64.71	**60.00**	10.90	100.00
gpt-turbo-3.5	82.35	10.72	**45.00**	42.11	3.33	**62.50**	80.00	**8.39**	100.00
gpt-4	76.17	13.07	80.00	40.00	2.76	88.24	100.00	15.74	100.00

Table 5: Human evaluation results. BUR and UER are from the first stage. Rating is from the second stage.

pilot annotation, the inter-annotator agreement of this stage is 68.97% Krippendorff's alpha, showing high agreement. As shown in Table 5, text-davinci-003 achieves the best performance on all the metrics in ACU-based scores, showing that human favors text-davinci-003 the most. It is worth noting that, Alpaca achieves fairness that is similar to that of gpt-turbo-3.5, demonstrating that small models can achieve similar fairness.

In the second stage, the annotators are asked to give a rating on the fairness of the summaries. This is more subjective because annotators can put different emphasis on different ACUs. This serves as another angle of fairness which gives more space to annotators' personal judgment on fairness. The Randolph's Kappa score is 0.41, demonstrating the high agreement across 4 annotators. As shown in Table 5, different from the ACU-based scores, *the overall ratings lean toward turbo models which align with the automatic metrics*.

6.4 Meta-Evaluation

We also use human evaluation to test the quality of proposed metrics. As shown in Figure 3, the correlation between the target distribution computed by BARTScore and human evaluation is around 0.91 when choosing 0.1 as the softmax temperature. This shows the high alignment of proposed metrics and human evaluation. Because BARTScore is not designed to show the proportion of entailment (For instance, 0.5 does not indicate that half of the summary can be inferred from the source.), we aligned BARTScore with human evaluation on absolute values by selecting 0.1 as temperature which enjoys a high Krippendorff's alpha with human evaluation.

6.5 Distribution across Values

We dissect the fairness of abstractive summaries generated by LLMs using the distribution across values. Figure 4 shows the distribution of males and females on Claritin. X-axis is the target distribution divided by source distribution

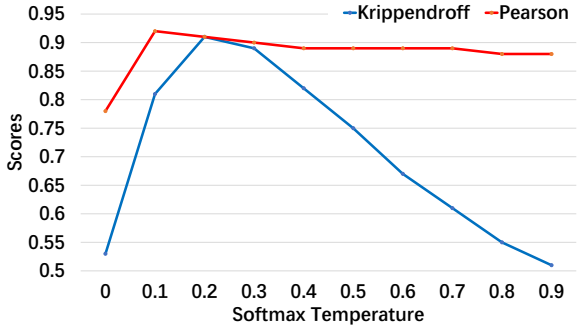

Figure 3: Relation between temperature and correlation scores on Claritin using gpt-turbo-3.5. X-axis is the softmax temperature of BARTScore. Y-axis is the Krippendorff's alpha and Pearson correlation coefficient with human evaluation. Pearson correlation coefficient is higher than Krippendorff's alpha because Pearson correlation coefficient only computes positive relations while Krippendorff's alpha requires the annotations to be the same.

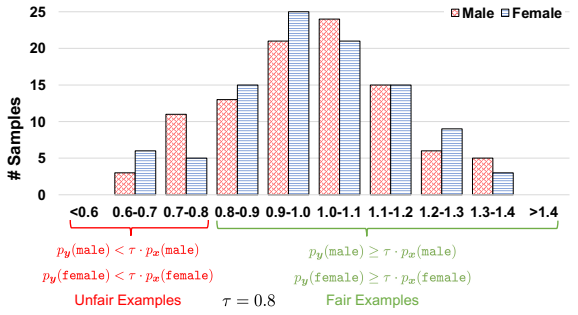

Figure 4: Distribution of Male and Female values in summaries generated by gpt-turbo-3.5 on Claritin.

$\boldsymbol{p_y}(v_k)/\boldsymbol{p_x}(v_k)$. As can be seen, about 20% of gpt-turbo-3.5 outputs contain underrepresented and unfair summaries for both female and male contents with the tolerance hyperparameter $\tau = 0.8$. *The proportion of unfair male/female tweets is similar, showing that the model generally performs well on Second Order Fairness (SOF is 0.17).*

6.6 Analysis on AUC

We compute the AUC score by aggregating BUR scores over different cutoff thresholds. This is use-

	N-gram↓	BERT↓	BART↓
SupremeCourt			
Alpaca	37.04	47.72	42.24
llama-2-chat-7b	25.14	36.48	41.75
llama-2-chat-13b	25.34	**31.59**	**40.72**
llama-2-chat-70b	**24.46**	34.23	41.19
text-bison@001	27.81	35.55	42.11
text-davinci-003	25.83	34.49	42.19
gpt-turbo-3.5	25.40	32.33	41.68
gpt-4	24.86	32.36	41.48
claude-instant-1	40.29	–	–
IQ2			
Alpaca	29.94	43.86	48.29
llama-2-chat-7b	25.59	40.76	47.66
llama-2-chat-13b	26.34	**40.13**	**46.85**
llama-2-chat-70b	25.63	49.06	47.80
text-bison@001	32.05	43.22	47.13
text-davinci-003	27.32	41.97	48.42
gpt-turbo-3.5	**24.00**	40.24	47.68
gpt-4	31.26	42.60	48.46
claude-instant-1	33.47	–	–

Table 6: AUC scores on SupremeCourt and IQ2.

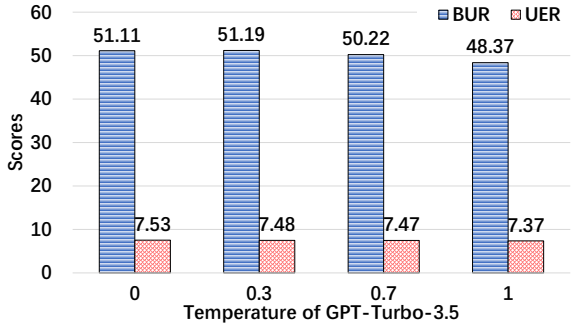

Figure 5: Effect of decoding temperature.

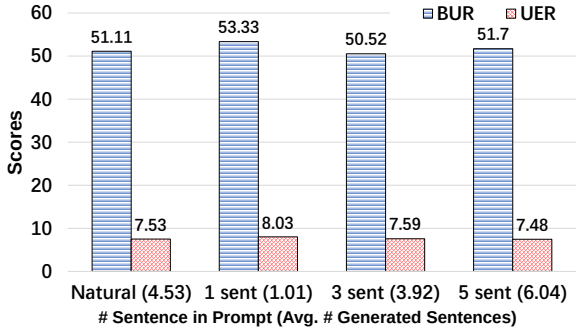

Figure 6: Effect of summary length in the number of sentences. Natural means not controlling the number of sentences, while 1/3/5 sent means that the model is prompted to generate a summary with 1/3/5 sentences.

ful for datasets containing more values $|\mathcal{V}|$ that require more flexible fairness. Thus, we compute the score for SupremeCourt and IQ2 datasets in Table 6. As shown in the table, LLaMA 2 obtains the best AUC score on SupremeCourt and IQ2. *This demonstrates that across different thresholds, LLaMA 2 obtains better performance than the GPT family on datasets with diverse values.* Claude model obtains a higher AUC score compared with other models as it consumes much longer data.

7 Improving Fairness of Abstractive Summarization

We propose three methods to improve the fairness of summaries by LLMs: decoding temperature, summary length, and fairness instruction.

7.1 Improving through Hyperparameters

Decoding Temperature. The temperature controls the proportion of novel words in outputs by modifying the sampling probability of tokens (Hinton et al., 2015). The model generates more novel words with higher temperatures (Appendix J). To test the effect of temperature towards fairness, we evaluate gpt-turbo-3.5 on various temperatures from $\{0, 0.3, 0.7, 1\}$ on Claritin. As shown in Figure 5, *the BUR and UER scores decrease significantly when the temperature rises*. This shows the model generates more diverse results when the temperature is higher, leading to more fairness in certain samples. However, the readability of the summary will be lower as some noise tokens are generated with higher temperatures. Thus, it is a trade-off between quality and fairness.

Summary Length. We analyze the influence of the length by using gpt-turbo-3.5 on Claritin to generate different numbers of sentences in summaries (Table 11). Figure 6 shows the BUR and UER scores for different lengths. Among them, 3 sentences demonstrate the best performance on fairness, while the 1 sentence shows the worst. *Therefore, medium length works best, and when there are too many/fewer sentences, balancing the value in summary contents is more difficult.*

7.2 Improving through Instructions

We modify the instruction by giving more information and definition of fairness in the prompt (Appendix H). Table 7 shows the performance of gpt-3.5-turbo on three datasets. *We found that adding the definition of fairness summaries can signifi-*

	BUR↓	UER↓	SOF↓
Claritin	51.11	7.53	**0.17**
+ Definition	**47.48**	**7.18**	0.98
Yelp	21.80	3.01	**1.25**
+ Definition	**21.73**	**2.98**	1.27
Election	72.67	5.00	**0.54**
+ Definition	**67.04**	**4.35**	1.03

Table 7: Effect of fairness instruction prompt. Adding the definition of fairness can significantly improve fairness.

cantly improve fairness. However, SOF increases after using definition instruction which shows that the models are biased towards one gender.

8 Related Work

Fairness and Bias in Natural Language Generation. People have been aware of social fairness and bias in natural language generation such as gender bias in machine translation and toxicity from prompt-based generation (Prates et al., 2020; Sheng et al., 2020; Gehman et al., 2020). Only a few efforts study the fairness of summarization (Carenini and Cheung, 2008; Shandilya et al., 2018; Jørgensen and Søgaard, 2021; Zhou and Tan, 2023; Chhikara et al., 2023; Liu et al., 2023b). Dash et al. (2019) proposed an analysis of whether the generated summaries fairly represent different social groups, such as gender or politics. Keswani and Celis (2021) proposed a framework that takes an existing text summarization algorithm as a black box to obtain a summary that is relatively more dialect-diverse. However, they are limited to unsupervised extractive approaches and restricted domains. Another relevant line of research is opinion summarization (Carenini et al., 2013; Gerani et al., 2014; Bražinskas et al., 2019, 2020b; Bhaskar et al., 2022; Iso et al., 2021), yet existing work mainly focused on few-shot learning instead of fairness as collecting reference summaries can be expensive. Therefore, our goal is to develop new metrics and datasets for fair abstractive summarization over diverse perspectives and values. Feng et al. (2023) and Santurkar et al. (2023) examined political biases in LMs stemming from training data and model persona, whereas our proposed benchmark assesses the ability of LLMs to generate fair abstractive summarizations.

Summarization Evaluation. Recent research reveals that the prior metrics such as ROUGE (Lin, 2004) and its variations (Rankel et al., 2013) cannot reliably evaluate LLM-generated summaries (Fabbri et al., 2021; Pillutla et al., 2021; Tam et al., 2022; Goyal et al., 2022b; Zhang et al., 2023; Chang et al., 2023a). New metrics of other dimensions have been proposed such as factuality (Kryściński et al., 2019) and controllability (Zhang et al., 2022). Liu et al. (2022, 2023a) propose a two-stage metric that predicts the Atom Content Unit (ACU) for the summary first and checks the proportion of the ACUs that can be inferred from the source text. Olabisi et al. (2022) proposed a benchmark for measuring the ability to represent salient as well as diverse perspectives. However, their evaluation metrics are unable to assess the distribution of summaries in abstractive summarization tasks.

Evaluation of LLMs. Recently, more attention has been attracted to the evaluation of LLMs (Hendrycks et al., 2020; Liang et al., 2022; Srivastava et al., 2022; Workshop et al., 2022; Li et al., 2023; Zheng et al., 2023; Chen et al., 2023). Some work evaluates the zero-shot ability (Qin et al., 2023), and multitasking, multilingual, and multimodal abilities (Bang et al., 2023), while others focus on the alignment and safety issues (Perez et al., 2022; Wang et al., 2023; Schulhoff et al., 2023). By contrast, we probe the fairness of LLMs by benchmarking fair abstractive summarization.

9 Conclusion and Future Work

In this paper, we systematically investigate fair abstractive summarization. Results of our metrics and human evaluation show that neither humans nor LLMs could maintain the fairness of summaries. Further analysis shows that prompt engineering and careful choosing of the temperature of LLMs can significantly improve the performance of fairness.

Future work includes fine-tuning small or large language models to improve the fairness of abstractive summarization, extending metrics to other types of bias, and developing datasets that apply the concept of fairness in other generation tasks.

Acknowledgment

We thank Greg Durrett, Shuyang Cao, Ranran Haoran Zhang, Sarkar Snigdha Sarathi Das, Fang Geng, Xi Li, and Kai Huang for the valuable discussions and comments. We also would like to thank the anonymous reviewers for their helpful comments.

Limitations

Our scope of the project is to focus on model fairness instead of bias in the dataset. We assume that the fairness of the training data of the models is unknown and only discuss the fairness of the generated summarization results. We believe we have covered a variety of domains and attributes in many socially impactful applications where fairness is critical, while our framework of evaluation can also be extended to other domains and attributes. In this paper, all datasets in PERSPECTIVESUMM are pre-cleaned by their respective authors, so we assume there are no invalid samples. However, we acknowledge the possibility of invalid samples.

Ethics Statement

PERSPECTIVESUMM contains six datasets that cover the typical domains and attributes that could cause fairness issues. We do not mean that these domains are the only ones that need attention for fairness. Due to the characteristics of the experiment, we pick some of the diverse domains and attributes to form PERSPECTIVESUMM. Second, we define the values for each attribute in a convenient and efficient way. It is possible that some values do not cover all possibilities, such as classifying gender in the Claritin dataset into female and male. Again, we use this dataset as the basic material for experiments, not meaning the source of PERSPECTIVESUMM is exactly aligned with social norms. Third, as indicated in the paper, our definition of fairness is flexible, and not restricted to ratio fairness. We conduct experiments using this criterion because it is one of the definitions that can reflect the fairness issue and enhance our conclusions. It does not show that ratio fairness is more important than other ones. In fact, our definition and code support various types of fairness, including but not restricted to ratio fairness, and equal fairness. Forth, we use the results to measure the fairness of the models, but this does not mean that our scores are the only metrics to judge the fairness of the summaries. Last but not least, the formation of PERSPECTIVESUMM may not align with real-world distribution due to the preprocessing algorithm. For Claritin, and Election, we randomly sample a certain number of twits to form one sample, however, these twits may not be able to reflect the real-world distribution. For Yelp and Amazon, we randomly sampled eight reviews which may also be different from real-world data. Also, for sentiment in the Yelp dataset, we use the NLTK tool to predict the sentiment of each review which may lead to misclassifications, though NLTK sentiment classification can achieve a high score. For dialogue datasets IQ2 and Oxford Debates, we segment the source into chunks due to the lengthy input. This may hurt the integrity of origin dialogue, leading to a difference between PERSPECTIVESUMM and original complete dialogue in these datasets.

References

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

Anthropic. 2023. Model card and evaluations for claude models.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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

Dan Biddle. 2017. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

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

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Few-shot learning for opinion summarization. *arXiv preprint arXiv:2004.14884*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Giuseppe Carenini and Jackie C. K. Cheung. 2008. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2013. Multi-document summarization of evaluative text. *Computational Intelligence*, 29(4):545–576.

Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. 2018. Fair and diverse dpp-based data summarization. In *International Conference on Machine Learning*, pages 716–725. PMLR.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023a. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023b. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

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

Garima Chhikara, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2023. Fairness for both readers and authors: Evaluating summaries of user generated content. In *SIGIR*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

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. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

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

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

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kung-Hsiang Huang, Philippe Laban, Alexander R. Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023a. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023b. Examining bias in opinion summarisation through the perspective of opinion diversity. *arXiv preprint arXiv:2306.04424*.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2021. Comparative opinion summarization via collaborative decoding. *arXiv preprint arXiv:2110.07520*.

Anna Jørgensen and Anders Søgaard. 2021. Evaluation of summarization systems across gender, age, and race. *arXiv preprint arXiv:2110.04384*.

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

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. *GitHub repository*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*.

Yixin Liu, Alexander R Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023a. Towards interpretable and efficient automatic reference-based summarization evaluation. *arXiv preprint arXiv:2303.03608*.

Yuhan Liu, Shangbin Feng, Xiaochuang Han, Vidhisha Balachandran, Chan Young Park, Sachin Kumar, and Yulia Tsvetkov. 2023b. What constitutes a faithful summary? preserving author perspectives in news summarization. *arXiv preprint arXiv:2311.09741*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

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.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Liu Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. Ignore this title and hackaprompt: Exposing systemic vulnerabilities of llms through a global scale prompt hacking competition. *arXiv preprint arXiv:2311.16119*.

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, Premkumar Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the factual consistency of large language models through summarization. *arXiv preprint arXiv:2211.08412*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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

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

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2022. Macsum: Controllable summarization with mixed attributes. *arXiv preprint arXiv:2211.05041*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Karen Zhou and Chenhao Tan. 2023. Characterizing political bias in automatic summaries: A case study of trump and biden. *arXiv preprint arXiv:2305.02321*.

A Dataset Construction

Details of these six datasets are listed as follows:

Claritin (Shandilya et al., 2018) contains tweets about the effects of the drug Claritin. Each tweet is annotated with the gender of the user who posted it. We use this dataset to evaluate gender fairness, where a = gender and $\mathcal{V} = \{\text{male}, \text{female}\}$. We randomly sample a number of tweets written by males and females with a certain ratio between them and then combine them together as one input source. This procedure is repeated multiple times with a diverse number of tweets and male ratios to form the whole dataset.

US Election (Shandilya et al., 2018) contains tweets posted during the 2016 US Presidential election. Each tweet is annotated as supporting or attacking one of the presidential candidates or neutral or attacking both. We use this dataset to evaluate politics fairness across parties, where a = politics and $\mathcal{V} = \{\text{pro-rep}, \text{pro-dem}, \text{neu}\}$. Similar to the Claritin dataset, we form the final dataset as a mixture of sampled tweets of certain ratios of each value in $\mathcal{V}$.

Amazon and **Yelp** are two datasets from the FewSum dataset (Bražinskas et al., 2020a) containing product and business reviews. Each sample in

these two datasets consists of multiple user-written reviews. Among these two datasets, FewSum provides reference summaries for 32 products on Amazon and 70 businesses on Yelp. We use the Amazon dataset to conduct opinion fairness analysis, where a = ratings, and $\mathcal{V} = \{1, 2, 3, 4, 5\}$ shows the ratings of each reviewer. We use the Yelp dataset to conduct sentiment fairness analysis, where a = sentiment, and $\mathcal{V} = \{\text{pos}, \text{neu}, \text{neg}\}$ displays the sentiment of reviews. We use NLTK (Loper and Bird, 2002) to produce sentiment for source text in Yelp because the dataset does not provide the sentiment of source reviews. We also use the reference summaries of the two datasets to further analyze the fairness of human-written summaries (Section 6.2).

SupremeCourt (Danescu-Niculescu-Mizil et al., 2012) contains a collection of conversations from the U.S. Supreme Court Oral Arguments from 204 cases involving 11 Justices and 311 other participants. We use SupremeCourt to analyze the fairness of speakers in court transcripts, where a = speakers, $\mathcal{V}$ are the names of all participants. For the input to the models, we truncate each transcript into shorter segments of k tokens to ensure the texts do not exceed the length limit of the models. We also include the whole text of each sample for the experiments of claude-instant-1.

Intelligence Squared Debate dataset (IQ2) (Zhang et al., 2016) collects public debates that follow the Oxford style and are recorded live. In each debate, teams of 2 to 3 experts debate for or against a motion and attempt to sway the audience to take their position. Each debate also has a moderator. We use the IQ2 dataset to analyze the fairness of speakers in debates, where a = speakers, $\mathcal{V}$ contains the names of all participants. We also truncate each transcript of the debate into shorter segments of k tokens. We include the whole text of each sample for the experiments of claude-100k.

It is worth noting that, two dialogue datasets contain more than one attribute, which emphasizes that people can have diverse perspectives of fairness on the same input source text.

B Dataset Example

Table 8 shows some examples in PERSPECTIVESUMM.

C Quality of the Metrics

We conduct an oracle test to evaluate our metric quality. To this end, we create extreme synthetic examples by biased/balanced sampling to test the upper/lower bound of BUR and UER scores. We randomly choose 100 examples from Claritin, and we replace the original summary with some tweets sampled from the source as the pseudo-summary. For biased sampling, we sample 5 male tweets from the source as the pseudo-summary, which should have high BUR and UER scores. For balanced sampling, we sample 5 tweets from the source as the pseudo-summary so that the target distribution is the same as the source, e.g., when the male ratio is 20%, we sample 1 male tweet and 4 female tweets from 10 source tweets to maintain the proportion. Although this is not strictly fair, because the length of each tweet varies, it should enjoy a very low BUR and UER scores.

As shown in Figure 7, The difference is significant between scores of bias/balanced samplings for all three metrics. Among the three scores, N-gramScore gives the lowest BUR/UER scores in biased sampling. This is because N-gramScore decomposes the summary of sampled male tweets into tokens during matching. Although the entire summary comes from male tweets, many tokens appear in female tweets in the source. This decreases the BUR/UER. In contrast, the other two metrics are based on semantic similarities.

Besides, we pick $k = 1$ for N-gramScore because using $k > 1$ makes the metrics meaningless. During our experimentation with higher values of k ($k = 2, k = 3$), we observed that in over 80% of samples, the overlap between the source and target summaries was zero. This is because our abstractive summarization tasks, particularly for shorter summarizations, rarely produce 2-gram or 3-gram overlaps. Consequently, using $k > 1$ makes the metrics meaningless as the fairness of the generated summary will equal generating an empty string. This underscores the necessity for our embedding-based metrics in such scenarios.

D Correlation between Metrics

We conduct an analysis to calculate the correlation between metrics, and the results are summarized in Table 9. As can be seen, BUR has a high correlation with UER and AUC which is in line with our assertion: BUR provides a binary indication of fairness, while UER and AUC offer detailed infor-

Claritin	@dararoseee have you seen the 1st one? word, i thought i'd avoided them, but i had a sneezing fit earlier this week. claritin on deck. ‖ Ima have to dope myself up on dis Zyrtec, Benadryl, r Claritin. ‖ Buttttttt I have some cute things, Claritin, a tank full of gas. So I guess I'll survive. ‖ @brooke_opland there my babies omg can't live without my Claritin ... *truncated contents* ... @nancyholtzman i heard that allergy medicine (antihistamine) can also dry up milk, depending on strength, claritin ok but.. @Jeannette108 ‖ It has taken at least 3 hours for the Claritin to kick in and make my nose slow from a running faucet, to just a leaky one. ‖ @thebrokenplate i just hope it IS just allergies. Is your throat sore too? Took claritin last nt but now have to wait 24 hrs to take again ‖ Dudes, I think the Claritin is making me...peculiar. I mean, moreso than usual.",
Election	Very strong case for Hillary by the Times, with context and perspective that's been hard to find this race. https://t.co/n8mJPZEjEf ‖ Gold Star families have made sacrifices most of us can't even fathom. They deserve our respect and our thanks. https://t.co/c9gDHudBjt ‖ Michael Bloomberg knows Donald Trump. And he's begging us to elect Hillary Clinton. Trump is dangerous. #ImWithHer https://t.co/JWC3rAb8Td ‖ Happy birthday to this future president. https://t.co/JT3HiBjYdj ... *truncated contents* ... ‖ This election is a choice between an economy that benefits everyone or an economy that benefits...Donald Trump. https://t.co/PEHnJDdiLq ‖ We can take on the threat of climate change and make America a clean energy superpower. Or, we can do nothing. https://t.co/JlYmN61epB ‖ Weird, I can't find anything on @Project_Veritas videos on @nytimes or @washingtonpost. It's almost like they're deliberately ignoring them.",
Yelp	I was pleasantly surprised with my meal . The burger and fries were seasoned to perfection . The willow bar dessert was mouth watering . The prices weren 't bad and the staff was friendly and helpful . We will be back ‖ Stopped there just for a drink and a little something to eat . We loved the wine list and the food . The menu had interesting selections and everything was really beautifully prepared and delicious . Worth a trip . ‖ Fantastic food . We had the Butternut squash ravs and hubby had the pork chop . Amazing . Drinks were great . Waitstaff cool and in the game . Can 't go wrong here ! ! ... *truncated contents* ... ‖ Easter dinner did not disappoint . The filet , and crab cakes are fantastic . Only wish we had bread served with dinner . Great job ! Service was great as usual . ‖ Great Dining experience . 730 reservations on a Saturday night . Seated promptly , waiter was attentive and informative . It was our first visit here so we followed his suggestions and it was spot on . Food was very good and atmosphere was great . A very good dining experience ." ,
Amazon	I know that the point of these bags is for them to be for small snack items , but they were too small for most of what I wanted to use them for . I think the normal sandwich sized bags work just as well . ‖ I used these for travel sized items being I hate spillage in my suitcase . I store small items such as loose herbs in them for immediate use . ... *truncated contents* ... ‖ THESE BAGS ARE THE BES , WE USE THEN IN EVERY WAY POSSIBLE AT HOME AND WHEN THE KIDS HAVE TO GO TO SCHOOL . ‖ I love snack size bags and am happy to find them on Amazon . These bags are perfect for packing snacks on the go without using bigger sandwich bags . I also use these smaller bags to organize office and hair supplies . ‖ I got 6 total boxes . I love the ziploc ones because of the double seal . The bags are reusable for about 5 times before they get nasty ." ,
SupremeCourt	**JUSTICE STEVENS** : We will now hear argument in the Cherokee Nation against Thompson and Thompson against the Cherokee Nation. Mr. Miller. **MR. MILLER** : Justice Stevens, and may it please the Court: These two contract cases concern whether the Government is liable in money damages under the Contract Disputes Act and section 110 of the Indian Self-Determination Act when the Secretary fails to fully pay a contract price for the – **JUSTICE O'CONNOR** : Would you mind explaining to us how these two cases relate? The Court of Appeals for the Federal Circuit decision went one way and the Tenth Circuit went another. And are the claims at all overlapping? How are they differentiated? **MR. MILLER** : No, Justice O'Connor. They're – they're not overlapping. T ... *truncated contents* ... **MR. MILLER** : Your Honor, we – we do not believe that that – that should be the outcome. That would advantage the contractors that came forward and not take account of the entire situation. We think the global situation has to be looked at. The total amount of the contracts that were not paid in fiscal year 1994 –
IQ2	**Bob Costas** : . . . And now I'd like to introduce Robert Rosenkranz, who is the chairman of the Rosenkranz Foundation, and the sponsor of Intelligence Squared, who will frame tonight's debate. Bob? This is Bob. **Bob Costas** : Thank you again, Bob. So this is the sixth debate of the second Intelligence Squared US Series. ... *truncated contents* ... **Robert Rosenkranz** : Well thank you very much. And, uh, uh, on behalf of, uh, Dana Wolfe, our executive producer and myself, uh, I'm just, uh, thrilled to welcome you. When we scheduled this event some, uh, five months ago, we had no idea it would be so timely. Just in the past month the, uh, Mitchell Report was released, naming some eighty eight Major League Baseball players alleged to have used steroids and, uh, uh, other drugs. Roger Clemens' denials have been heard in 60 Minutes and were front page news in Sunday's New York Times. Uh, record breaking sprinter Marion Jones was sentenced to six weeks in prison-, or six months, I ... *truncated contents* ...

Table 8: Six samples from PERSPECTIVESUMM. One sample for each dataset. Reviews are shuffled and concatenated by "‖". The names of the speakers in SupremeCourt and IQ2 datasets are in **bold**.

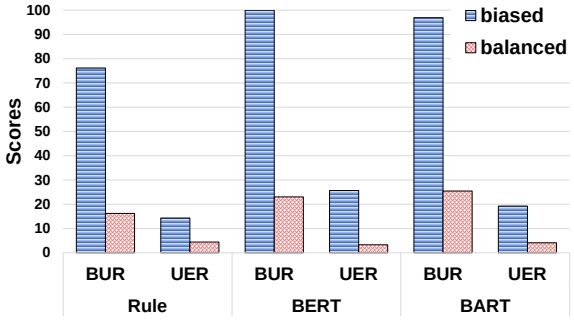

Figure 7: Extreme synthetic cases to validate the quality of metrics.

	BUR	UER	AUC	SOF
BUR	1.00	0.56	0.44	0.08
UER	0.56	1.00	0.27	0.43
AUC	0.44	0.27	1.00	0.06
SOF	0.08	0.43	0.06	1.00

Table 9: Pearson correlation between metrics on Claritin dataset.

mation across different UER set S_{UER} or thresholds τ, respectively. Similarly, the strong correlation between SOF and UER supports our design rationale, as UER represents the mean of the UER set, while SOF reflects the coherence of this set. When UER is high, it suggests that the elements in the UER set are likely dispersed along the axis.

E Human Evaluation

We introduce the design principles of human evaluation, meta-evaluation, and evaluation guideline details.

E.1 Design of Human Evaluation

To clearly measure the performance of the models and validate our automatic metrics, we design a two-stage human evaluation by asking four annotators with English expertise and experience in summarization to manually label 20 examples in Claritin, 20 in Yelp, and 10 in IQ2.

The two annotation tasks include Sentence Fact Identification and Summary Fairness Identification. Sentence Fact Identification produces a score using step-by-step instruction because the annotators need to read and identify facts and add them together to produce target value distribution. On the contrary, Summary Fairness Identification gives an end-to-end score that directly labels the rating of fairness after reading the definitions. We would like to use these two types of explicit and implicit computation of fairness to make human evaluation more comprehensive and diverse.

ACU-based Score. The first task is Sentence Fact Identification. This task asks annotators to label the target value distribution of the summaries. Annotators are firstly asked to split the summary sentence into facts where each fact is the atom semantic unit which is defined as the smallest unit that holds semantic information. Next, the source text is split into different values, and the annotators need to identify where each atom semantic unit comes from, which means the input perspectives that can fully infer (entail) this fact. Compared with direct operation on each sentence, this operation enjoys two advantages. First, each fact contains the same amount of information so that they can be added together without defining weights. Second, it is easier for the annotator to judge where it is from. After obtaining the human-annotated target distribution based on the ratio of atom semantic units, we run the automatic program to compute the BUR and UER metrics. We call this an ACU-based human evaluation score.

Rating-based Score. The second task is Summary Fairness Identification. In this stage, the annotators are asked to judge the fairness of each value. Different from stage one's computation, annotators need to give an overall judgment of the level of fairness. We assign 5 ratings, from 1 to 5, where 1 indicates only representing 0%-20% of the input distribution and 5 indicates over 80%. Thus, only 5 indicates a fair summary, while 1-4 is all unfair with different levels. We call this a rating-based human evaluation score. More details of annotation guidelines are available in Appendix E.2.

E.2 Guidelines of Human Evaluation

We list part of the guidelines for annotators, including instructions for Sentence Fact Identification and Summary Fairness Identification. We also show our annotation interface.

Sentence Fact Identification For this task, the annotators need to complete three steps:

- **Decompose summary into facts**. Please decompose the generated summary into facts. Facts in our project are countable (e.g. we can say sentence 1 contains 2 facts), usually, a longer sentence contains more facts. We provide a list of ACUs (Atomic Content Units)

for the annotator to refer to, however, these ACUs may contain errors and inaccurate expressions. Annotators can decide how many units are there and what they are according to our own understanding of facts —- the smallest semantic unit. As Figure 9 shows, annotators can modify ACUs in the marked column (add, delete, or modify) to accurately show the semantic units of the sentences.

- **Identify the source of the facts**. After getting all the facts in the summary, the next step is to identify where each fact comes from. "Comes from" here means the input source text that can fully infer (entail) this fact. We need to carefully read the input of different values, and then annotate which value the fact is generated from. There are two main cases, 1) the fact can be inferred from a single value, and 2) it is not possible to be inferred from any values (hallucination). For 1), add a counter for that value(s) (note that probably more than one value can infer this fact). For 2), add a counter for hallucination.

- **Count the facts for all values**. After annotating the source of each fact, the program will automatically generate the count for each value. Although this step is done automatically, the annotator needs to check it to ensure fact annotation and the calculation are correct.

Summary Fairness Identification Summary Fairness Identification only contains one step:

- **Rate the level of fairness**. You can refer to the ratio of facts in the spreadsheet. Give a score to rate if each value is underrepresented. We have 5 ratings for being underrepresented, if you find the ratio of one value in summary is larger or equal to 80% of that value in source, the rate is 5. Overall, rate it according to the proportion: 5: >80%, 4: 60%-80%, 3: 40%-60%, 2: 20%-40%, 1: 0%-20%

Annotation Interface Figure 8 shows an example of the SummVis interface. Annotators can click on the words with underlines to see the source of the tokens to improve the annotation speed and accuracy. Finally, the results of annotations will be collected in a spreadsheet as shown in Figure 9 for further computation.

F Model Details

We compare the fairness performance of 5 models that are representative LLMs with the state-of-the-art performance on summarization tasks. For all models, we use a zero-shot prompt setting because LLMs have a strong capability for summarization (Goyal et al., 2022a; Zhang et al., 2023) and we do not want the reference summaries to influence the judgment of the LLMs themselves. The LLMs studied are listed as follows:

GPT We use three models by OpenAI: two variants of GPT-3.5 (**text-davinci-003** and **gpt-3.5-turbo**) (Brown et al., 2020; Ouyang et al., 2022) and **gpt-4** (OpenAI, 2023). The details regarding the training process of these models have not been published, but the authors mention that GPT-4 is trained on "publicly available data (such as internet data) and data licensed from third-party providers". Therefore, it is possible that these models are trained on summarization datasets. For parameters, we set the temperature to 0 and the maximum length to 512.

LLaMA 2 (Touvron et al., 2023b) is an open-sourced pretrained and fine-tuned LLM ranging in scale from 7 billion to 70 billion parameters. We use **llama-2-7B-chat**, **llama-2-13B-chat**, and **llama2-70B-chat** which are optimized for instruction following and dialogue use cases. We set the temperature to 0 and the maximum length to 512.

Alpaca (Taori et al., 2023) is a model fine-tuned from the LLaMA model (Touvron et al., 2023a) on 52K instruction-following demonstrations[2]. The training data for LLaMA is acquired from seven sources, and it does not include summarization datasets. However, the training data for Alpaca includes examples of text summarization in various domains, including news articles and scientific papers. For parameters, we use a temperature of 0.1 and a context length (the maximum number of tokens from both input and output) of 2048.

PaLM 2 (Anil et al., 2023) is a close-sourced LLM optimized for multilingual and reasoning tasks. It is also a Transformer-based model trained using a mixture of objectives. We directly use **text-bison@001** model provided by Google Vertex API[3] to run all experiments. We set the temperature

[2] https://github.com/tatsu-lab/stanford_alpaca

[3] https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text?hl=zh-cn

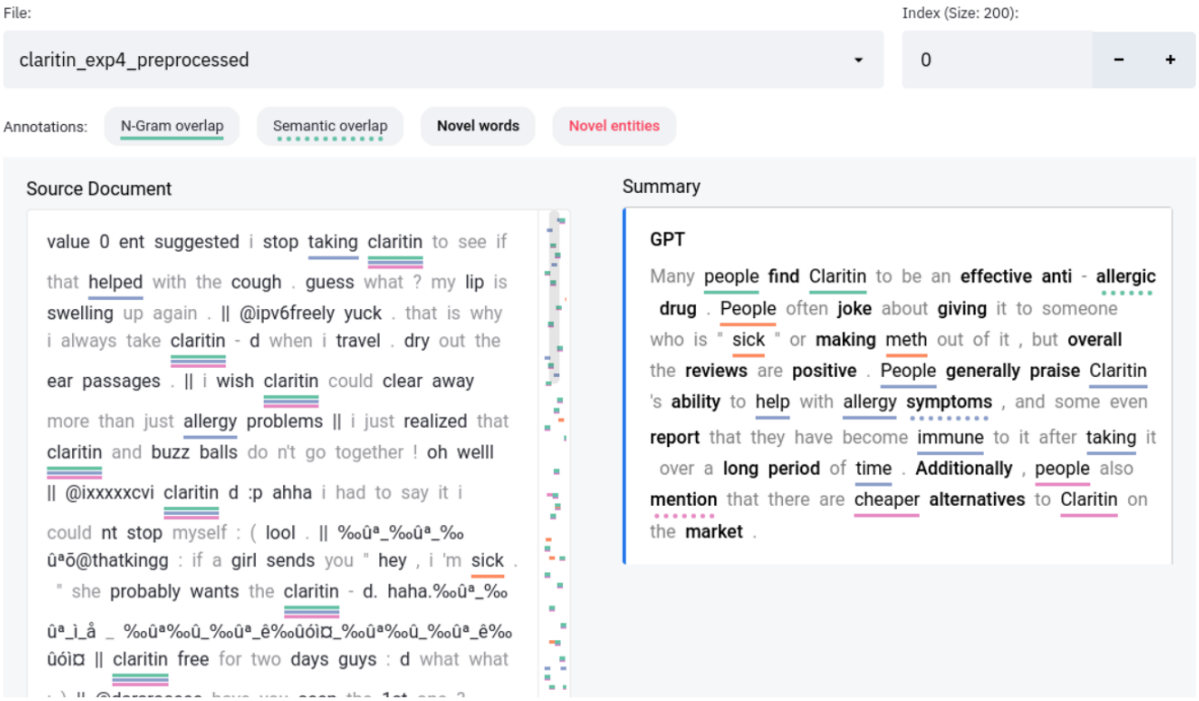

Figure 8: A sample for SummVis annotation interface. There are 4 types of tokens labeled by SummVis: N-gram Overlap, Semantic Overlap, Novel words, and Novel entities.

acu_id	acu	value 0	value 1
0	Many people find Claritin to be an effective anti-allergic drug.	0.5	0.5
1	Claritin is also known as Benadryl.		
0	People often joke about giving it to someone who is "sick"	1	0
1	People often joke about making meth out of it	1	0
2	the overall reviews are positive	0.5	0.5
0	People generally praise Claritin's ability		
1	People generally praise Claritin's ability to help with allergy symptoms	0.5	0.5
2	Some report that they have become immune to Claritin		
3	People report that they have become immune to Claritin after taking it		
4	People report that they have become immune to Claritin after taking it ov	1	0
0	People also mention that there are cheaper alternatives to Claritin	1	0
1	there are cheaper alternatives to Claritin on the market	1	0
		6.5	1.5

Figure 9: A sample for annotation spreadsheet. The annotator can remove the ACUs (marked in red), edit the existing ACUs (marked in green), or add new ACUs during annotation. Then the annotators give scores to each value of the attribute for the ACUs.

to 0 and the maximum length to 512.

Claude is an Anthropic's model.[4] We use **claude-instant-1**, and the details regarding the training process of this model have not been published. Claude can help with various tasks, such as summarization, search, creative and collaborative writing, Q&A, and coding. claude-instant-1 is one of the claude models which can consume 100k tokens.[5] All our experiments use their default hyper-parameters.

G Result Details

Table 10 shows the details of each score, including N-gramScore, BERTScore, BARTScore, and their average scores.

H Prompt Templates

For each dataset in PERSPECTIVESUMM, we design a prompt for LLMs. For controlling sentences and adding fair prompts, we design two prompts as well. The prompt templates are listed in Table 11.

I Model Comparison

The Claritin dataset consists of different male ratios and number of tweets. We compare the performance of gpt-turbo-3.5 and gpt-4 in detail by decomposing the results into each (male ratio, number of tweets) pair. Figure 10 compares the gpt-turbo-3.5 and gpt-4 models in terms of the number of fairness samples on the Claritin dataset. The value t in a cell means that the gpt-turbo-3.5 has t more fair samples than gpt-4. As shown in the table, gpt-turbo-3.5 outperforms gpt-4 when the male ratio is 0.5 and the number of tweets is 10 while gpt-4 is better in the other cases. This shows that gpt-4 is better at processing longer tweets and the samples where the male ratio is not equal.

Dataset Constitution Figure 10 shows the number of error samples for each combination of male ratio and number of tweets. As can be seen, the fairness of the summaries decreases when the male ratio is not 0.5 where the ratio of male and female are the same. Besides, there is no significant difference between different numbers of tweets.

[4] https://docs.anthropic.com/claude/reference/selecting-a-model

[5] https://www.anthropic.com/index/introducing-claude

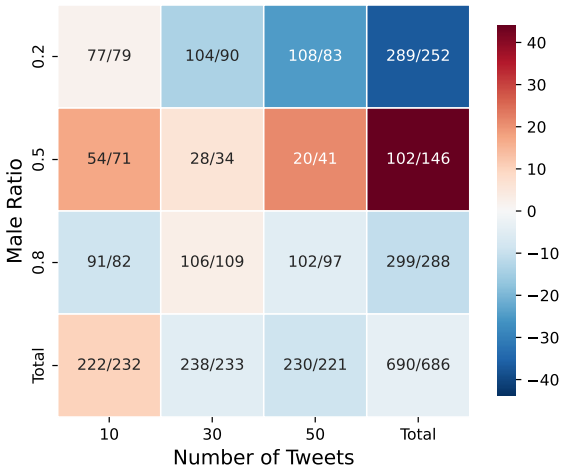

Figure 10: Error comparison between gpt-turbo-3.5 and gpt-4 on Claritin. Each row shows a different male ratio while each column is a different number of tweets per sample. Each number in the cell is the number of the unfair samples generated by gpt-turbo-3.5/gpt-4. The colors of the cells become redder/bluer when gpt-turbo-3.5 produces more/less fair samples than gpt-4.

J More Example Outputs

Table 14 shows two cases of model-generated summaries. Both cases turn the unfairly generated summaries into fair ones by adding fair instructions (Case 1) and by changing temperature (Case 2).

Table 15 shows two samples of reference summaries on Yelp and Amazon datasets. In sample one, the reference summary ignores the negative reviews which is important for the users to gain a comprehensive understanding of the product. For sample two, the summary overly emphasizes the negative aspects while disregarding the review that awarded a five or four-star rating. In both examples, our metrics successfully capture that these samples are unfair, as well as underrepresenting which values.

K Analysis on Fairness Improvement

We conduct a human evaluation task assessing the factuality, readability, fluency, and coherence of the proposed approaches, incorporating diverse lengths, temperatures, and fair prompts.

Specifically, we ask annotators to annotate 20 summaries from Claritin dataset generated by gpt-turbo-3.5 with various settings. For each summary, the annotator needs to annotate three metrics: For factuality, annotate one if any factuality errors are found, otherwise zero. For readability & fluency, and coherence, the annotators need to rate from one to five to indicate the level of performance (five

	Claritin		US Election		Yelp		Amazon		SupremeCourt		IQ2	
	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓	BUR↓	UER↓
Alpaca★ (7B)												
N-gramScore	51.12	7.63	82.56	7.01	34.47	3.71	74.27	5.11	70.77	5.41	82.14	4.37
BERTScore	70.85	12.17	83.27	6.83	38.73	5.52	76.20	6.89	78.46	6.07	85.94	6.94
BARTScore	73.84	11.13	94.31	9.72	58.93	6.11	91.80	9.48	89.23	7.09	90.78	10.32
Average	65.27	10.31	86.71	7.85	44.04	5.11	80.76	7.16	79.49	6.19	86.29	7.21
llama-2-chat★(7B)												
N-gramScore	47.56	7.11	57.48	3.64	16.67	2.71	61.87	3.94	95.35	4.03	73.05	3.43
BERTScore	67.63	10.70	84.37	6.28	47.60	3.09	73.47	3.95	96.84	4.42	87.26	5.72
BARTScore	73.78	9.46	94.07	10.14	61.00	4.97	92.67	7.39	100.00	6.31	92.93	8.92
Average	62.99	9.09	**78.64**	6.69	41.76	3.59	76.00	5.09	97.40	4.92	84.41	6.02
llama-2-chat★(13B)												
N-gramScore	47.04	7.01	60.07	3.89	21.80	3.02	64.33	4.23	95.34	4.13	76.21	3.68
BERTScore	68.37	10.98	82.59	6.23	29.40	2.81	64.60	3.95	91.88	3.91	89.30	6.45
BARTScore	73.78	10.07	95.19	10.26	57.40	4.76	91.60	7.16	99.25	5.70	92.61	9.85
Average	63.06	9.35	79.28	6.79	36.20	3.53	73.51	5.11	95.49	4.58	86.04	6.66
llama-2-chat★(70B)												
N-gramScore	42.74	6.60	66.00	4.16	22.33	3.26	64.60	4.43	95.19	3.96	75.23	3.59
BERTScore	66.00	10.70	81.63	6.07	30.00	0.03	67.87	4.63	95.79	4.21	87.68	6.26
BARTScore	75.85	12.47	91.70	8.99	58.80	5.13	91.60	7.82	99.85	6.05	94.23	10.07
Average	**61.53**	9.92	*79.78*	6.41	37.04	2.81	74.69	5.63	96.94	4.74	85.71	6.64
text-bison@001(N/A)												
N-gramScore	52.37	7.77	74.57	6.11	32.76	3.66	75.53	5.52	96.53	4.29	81.84	4.50
BERTScore	72.15	12.47	87.78	8.24	43.07	5.81	84.73	8.56	96.84	5.50	90.64	8.15
BARTScore	78.67	11.24	97.23	11.77	61.11	6.65	93.00	11.22	99.85	7.17	95.29	11.98
Average	67.73	10.49	86.53	8.71	45.65	5.37	84.42	8.43	97.74	5.65	89.26	8.21
text-davinci-003 (175B)												
N-gramScore	45.41	6.88	73.04	5.22	26.87	2.93	67.80	4.14	96.09	4.28	78.11	3.89
BERTScore	69.48	11.18	82.30	6.64	40.27	4.03	78.27	5.79	96.09	4.63	88.67	6.92
BARTScore	73.93	9.17	92.89	9.41	62.13	6.09	93.60	9.86	100.00	6.84	94.72	10.78
Average	<u>62.94</u>	**9.08**	82.74	7.09	43.09	4.35	79.89	6.60	97.39	5.25	<u>87.17</u>	7.20
gpt-turbo-3.5 (175B)												
N-gramScore	51.11	7.53	72.67	5.00	21.80	3.01	63.13	4.12	96.84	4.16	72.48	3.37
BERTScore	68.22	11.04	80.37	6.10	33.80	3.70	71.20	4.93	92.93	3.76	86.77	5.68
BARTScore	73.56	8.98	91.11	8.35	60.33	5.30	93.33	8.40	100.00	6.00	94.30	9.15
Average	64.30	<u>9.18</u>	<u>81.38</u>	**6.48**	**38.64**	<u>4.00</u>	<u>75.89</u>	<u>5.82</u>	**96.59**	<u>4.64</u>	**84.52**	**6.07**
gpt-4 (N/A)												
N-gramScore	50.81	7.51	65.78	4.38	20.27	2.50	59.87	3.88	95.34	4.12	79.45	3.87
BERTScore	70.44	11.96	78.81	6.22	36.60	3.21	72.00	4.39	95.04	3.71	88.88	6.96
BARTScore	77.85	10.34	95.19	10.36	61.40	5.46	92.47	8.20	100.00	5.87	94.79	10.00
Average	66.37	9.94	**79.93**	<u>6.99</u>	<u>39.42</u>	**3.72**	**74.78**	**5.49**	<u>96.79</u>	**4.57**	87.71	<u>6.94</u>

Table 10: Main results with our proposed metrics. BUR and UER are better with a lower score ↓. **Bold** indicates the best average score observed across of all models (consistent with the metrics used in Table 2), and <u>underline</u> indicates the second best. ★: Open source models.

indicates the best performance). Table 13 shows the average score of each summary in various settings.

As shown in the table, readability, fluency, and coherence decrease with temperature and length increasing. For factuality, experiments show that the proposed approaches do not significantly hurt factuality.

L Fairness Instruction on LLaMA

We conducted an experiment with the llama2-13b model, revealing that after adding the definition, fairness increased for Claritin dataset but decreased for Yelp and Election (Table 12). We speculate that the impact of adding a fairness prompt is linked to the instruction-following capability of LLMs. To validate our hypothesis, we conducted further qualitative analysis, finding that the results of LLaMA 2 contained noises after adding a fairness definition, such as repeating the question. For instance, redundant sentences like "The generated summary will be more fair..." are found in generated summaries.

Template for All Experiments on PERSPECTIVESUMM	
Claritin	Reviews about Claritin. 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:
US Election	Reviews about US Presidential Election. 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:
Amazon	Reviews about a product. 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:
Yelp	Reviews about a business. 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:
SupremeCourt	Dialogue of the Supreme Court oral arguments. Each turn of the dialogue is one line: {SOURCE} The summary of the dialogue is:
IQ2	Debates on certain topics. Each turn of the dialogue is one line: {SOURCE} The summary of the dialogue is:
Template for Analysis on PERSPECTIVESUMM	
Sentence Control	Summary it in {NUMBER} sentences.
Fair Instruction	{MALE_RATIO}% of the reviews are written by males and {1-MALE_RATIO}% written by females. They are mixed randomly in the source text. Please ensure the length of the male review in the summary is still {MALE_RATIO}% of the total length.

Table 11: Prompt template on PERSPECTIVESUMM. The first six templates are used for every sample in six datasets. The last two templates are used as additional prompts appended to the input for the use of analysis.

	BUR↓	UER↓	SOF↓
Claritin	47.04	7.01	0.09
+ Definition	**45.26**	**6.92**	**0.07**
Yelp	**21.80**	**3.02**	**1.53**
+ Definition	23.13	3.25	1.78
Election	**60.07**	**3.89**	0.94
+ Definition	68.44	4.36	**0.28**

Table 12: Effect of fairness instruction prompt.

model type	factuality↓	read. & flu.↑	coherence↑
sent=1	0.70	**4.65**	**4.75**
sent=3	0.55	4.60	4.60
natural	**0.40**	4.40	4.30
sent=5	0.60	2.90	3.70
temp=0	**0.40**	**4.40**	**4.30**
temp=0.3	0.45	4.35	4.20
temp=0.7	0.55	3.75	3.80
temp=1	**0.40**	4.10	3.70
w/o definition	0.40	4.40	**4.30**
w/ definition	**0.40**	**4.45**	3.80

Table 13: Human evaluation on factuality, readability & fluency, and coherence. Factuality scales from zero to one (zero is the best). The other two metrics scale from one to five, indicating the rating of the performance (five is the best).

Case 1: Influence of Instructions	
Source★	**Male Tweets**: 'mucinex and claritin. (@ cvs pharmacy) http://t.co/bju2jidt ‖ if a girl sends you "heyy, im sick" she probably wants that claritin-d', **Female Tweets**: '2 claritin 5 aspirins and an entire bag of cough drops later and i still feel like shit ‖ "íf says sheś sick, she probably wants the d.ćlaritin d?" - mom hahahaha i love my mom!!! ‖ my mom said i was grounded because iḿ threatening to overdose on claritin. ‖ i woke wit these allergies bout to pop sme claritin and get to wrk ‖ i took a claritin and motrin 30 minutes ago! ‖ i have claritin rt @juandelay: anybody got some oatmeal or benadryl?? im havin a allergic ‖ @jonathanrknight i prefer claritin, works better for me. ‖ thereś clear, and then thereś claritin clear'
w/o instruction	The reviews about Claritin, an anti-allergic drug, are mixed. Some people find it effective in relieving their allergies, while others mention using it as a joke or experiencing negative side effects. Overall, the reviews highlight different experiences and opinions regarding the use of Claritin.
w/ instruction	Claritin is a popular anti-allergic drug that seems to work well for many people. Some individuals use it to alleviate allergies and get through their day, while others joke about its potential effects. Overall, the reviews indicate that Claritin is effective in providing relief from allergy symptoms.
Case 2: Influence of Temperature	
Source★	**Male Tweets**: if a girl says she's sick. give her the claritin d. ‖ @bass_courtney: @21ccharlesworth oh my! yess i need some claritin -d.:p @shewantyourdick add her to the d list **Female Tweets**: @sawng-byrd28 yes you can overdose on claritin. any signs of muscle contractions, drowsiness, or headache? ‖ @kamandermainy: this sinus pressure is not the business i just got over mine.... don't u hate that? claritin d will clear it right up tho ‖ just took this claritin... no snoring for me tonight lol. ‖ in big trouble. forgot to buy claritin-d. may have to get a hotel for the night. #madhusbandwhenigethomeemptyhanded ‖ @4jinxremoving i'm already on claritin and nasonex!!! ‖ @so_freshgiraffe : actually, i'm allergic to cats. you know; red eyes, sneezing, hives. but i can't stay away from him. claritin helps (: ‖ smh it never fails every time i come home my allergies start actin up!!—but its all good cause im prepared this time! *pulls out claritin* ‖ @morethymeblog hi! good day today. took 1/2 of a claritin bcz allergies has been acting up on top of everything. my mom's b'day today! you?
Temp=0	People have found relief from allergies and sinus pressure using Claritin, with some even avoiding snoring at night. However, there are concerns about overdosing on the medication, which can cause muscle contractions, drowsiness, or headaches.
Temp=1	Some Twitter users mentioned needing Claritin to relieve allergies, particularly sneezing, red eyes, and sinus pressure. The drug appears to be effective in providing relief for these symptoms. It is mentioned that overdosing on Claritin is possible, which can lead to muscular contractions, drowsiness, and headache. Overall, many of the reviewers expressed satisfaction with Claritin's effectiveness in managing their allergies.

Table 14: Two case studies on Claritin. Predictions are generated by gpt-turbo-3.5. After adding fair instructions, the generated summary becomes fair (BUR becomes zero). After tuning the temperature from zero to two, the generated summary becomes fair (BUR becomes zero). ★ For the input to the model, male/female are not marked and all tweets are randomly shuffled; this table shows these gender values for the purpose of clarity.

Sample One from Yelp Dataset	
Source★	**Positive Reviews**: Well the atmosphere is excellent - especially in CU. The staff was all very friendly and attentive. Kudos to them! The food was a bit better than average but nothing to knock your socks off. A good place to go with family or friends.They have a cheese sauce that is a bit spicy but tasty! ‖ Great location and atmosphere. Authentic and Mexican American fare that was served in generous portions and DELICIOUS! The Guadalajara burrito was simple and SO flavorful. Hubs had Molcalambres and ranked them near the top of his fajita list. Not hoiity toity; not trying to be hip and trendy like Maize. Will come back whenever in town. ‖ Love this place! The waitstaff is great and friendly, they are very quick but also allow you to relax and not feel rushed. I go here in the summer with some friends and we just sit on the patio (great outdoor seating) and drink margaritas and enjoy the complimentary fresh chips and salsa. Also great food! ‖ They took a while to bring the food, my son ordered a chicken quesadilla that was very soggy and tasteless, the rest of the food was edible but far from good, place was ok, it's far from the best but if you hungry and desperate you can certainly go eat there. ‖ Good food at a great price, the restaurant has become a great lunch destination. The food is basic American-Mexican combinations (burrito, enchilada, tamale for example) with very quick service. Honestly have not been to a sit down restaurant with such speedy service during the lunch rush. ‖ Their food is decent, but the reason I'm writing this review is their healthy margarita. It's incredible and I'm not even usually a margarita fan! My husband ordered the regular margarita and we both agreed that the healthy was much better than the regular margarita. Go try it! ‖ Went here with a group to drink some margaritas and eat some chips and salsa. Wonderful service and friendly staff. I ordered a grande raspberry margarita and it was nice and strong. Chips were hot and delicious and salsa was good (wish it was spicier but that's a personal preference), **Negative Reviews** The first time I went here, I ordered a taco salad with beef, and the beef was pink. Not brown on the outside and pink on the inside, pink pink pink. The second time I went, my friend and I had the same burrito combination platter and both of us got sick the next day.
Reference Summary	This Mexican restaurant is a great place to get a margarita with chips and salsa. Their margaritas are very well done, and the chips and salsa are also good. The service is also pretty good, friendly and efficient. The rest of the food, however, is average at best. This restaurant is recommended mostly for having some drinks with friends.
Sample Two from Amazon Dataset	
Source★	**Rating=1** I'm 5 '4 'and this tank fits like a normal tank top, not any longer. I was trying to find a tank that would cover past my hips, so I could wear it with leggings. Don't order if you're expecting tunic length. **Rating=2** The only reason I'm rating this at two stars is because it is listed as a 'long' tank top and the photo even shows it going well past the models hips, however I'm short and the tank top is just a normal length. **Rating=3** The description say it long... NOT so it is average. That's why I purchased it because it said it was long. This is a basic tank.I washed it and it didn't warp but did shrink a little. Nothing to brag about. ‖ This shirt is OK if you are layering for sure. It is THIN and runs SMALL. I usually wear a small and read the reviews and ordered a Medium. It fits tight and is NOT long like in the picture. Glad I only purchased one. ‖ I usually get them someplace out but they no longer carry them. I thought I would give these a try. I received them fast, although I did order a brown and got a black (which I also needed a black anyway). They were a lot thinner than I like but they are okay. **Rating=4** The tank fit very well and was comfortbale to wear. The material was thinner than I expected, and I felt it was probably a little over priced. I've bought much higher quality tanks for $5 at a local store. ‖ I bought this tank to wear under shirts when it is colder out. I bought one in white and one in an aqua blue color. They are long enough that the color peeks out from under my tops. Looks cute. I do wish that the neck line was a bit higher cut to provide more modest coverage of my chest. **Rating=5** Every women should own one in every color. They wash well perfect under everything. Perfect alone. As I write I'm waiting on another of the same style to arrive. Just feels quality I don't know how else to explain it, but I'm sure you get it ladies!
Reference summary	Some Twitter users mentioned needing Claritin to relieve allergies, particularly sneezing, red eyes, and sinus pressure. The drug appears to be effective in providing relief for these symptoms. It is mentioned that overdosing on Claritin is possible, which can lead to muscular contractions, drowsiness, and headache. Overall, many of the reviewers expressed satisfaction with Claritin's effectiveness in managing their allergies.

Table 15: Two samples of unfair human-written reference summaries. Sample one is from Yelp dataset and ignores the negative review. Sample two is from Amazon dataset and ignores reviews with four or five stars. Underline words indicate the aspects that the reference summary ignores. ★ For the input to the model, sentiment/ratings are not marked and all reviews are randomly shuffled; this table shows these values for clarity.