Gender Bias in News Summarization: Measures, Pitfalls and Corpora

Anonymous ACL submission

Abstract

Summarization is an important application of large language models (LLMs). Most previous evaluation of summarization models has focused on their performance in content selection, faithfulness, grammaticality and coherence. However, it is well known that LLMs reproduce and reinforce harmful social biases. This raises the question: Do these biases affect model outputs in a relatively constrained setting like summarization?

To help answer this question, we first motivate and introduce a number of definitions for biased behaviours in summarization models, along with practical operationalizations. Since we find that biases inherent to input documents can confound bias analysis in summaries, we propose a method to generate input documents with carefully controlled demographic attributes. This allows us to study summarizer behavior in a controlled setting, while still working with realistic input documents.

Finally, we measure gender bias in English summaries generated by both purpose-built summarization models and general purpose chat models as a case study. We find content selection in single document summarization to be largely unaffected by gender bias, while hallucinations exhibit evidence of downstream biases in summarization.

1 Introduction

Pretrained large language models (LLMs) have increasingly found application across a wide variety of tasks, including summarization (Lewis et al., 2020; Zhang et al., 2020; Goyal et al., 2022). While such models often evaluate favourably especially in human judgement for content (Goyal et al., 2022), it is also well known that pretrained language models can often carry undesirable social *biases* (Dinan et al., 2020; Liang et al., 2022; Bommasani et al., 2021). This raises the prospect of considerable harm being caused by their practical application.

However, often these biases are studied in settings where model inputs are specifically crafted to help reveal social biases (Rudinger et al., 2018; Sheng et al., 2019; Parrish et al., 2022). Biases are also often observed in relatively unconstrained settings, such as dialog (Dinan et al., 2020) or the generation of persona descriptions (Cheng et al., 2023). While insights won in this way are highly valuable in understanding the potential negative impacts of LLMs, it is not always clear how these biases map to other applications.

Summarization in particular is a highly conditional task. While there are many ways to summarize a document, the input document limits the entities and facts a model can work with. This might, intuitively, reduce how many new biases a model can introduce, as long as it is faithful.

This leads us to ask: *How can we study bias in text summarization?* and *To which extent do current models exhibit biases when applied to text summarization?* We focus on gender bias in English language single document news summarization. Gender bias is a well-known issue in LLMs (Bolukbasi et al., 2016; Zhao et al., 2018; Dinan et al., 2020; Saunders and Byrne, 2020; Bartl et al., 2020; Honnavalli et al., 2022, among others), making it a useful phenomenon to develop fundamental methodology for studying bias in text summarization. We exclusively consider male and female identities, since it is a well studied group disparity and has grammatical indicators in many languages that are likely to be recognized by language models, leaving the extension to varied gender identities to future work. We select single document news summarization since it is a popular task (See et al., 2017; Narayan et al., 2018; Lewis et al., 2020; Zhang et al., 2020) which is performed well by current models (Goyal et al., 2022) and also a likely application of summarization models.

While an ideal evaluation would be conducted on naturally occurring data, we find that it is dif-

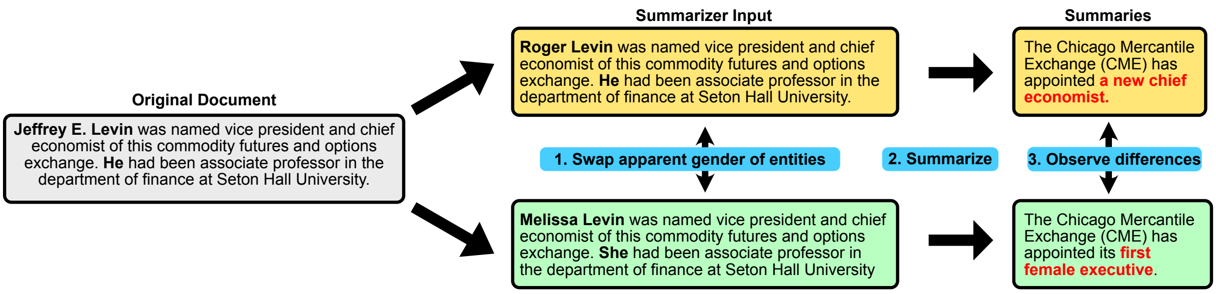

Figure 1: Schematic overview of our approach for summary gender bias evaluation with an example generated by BART XSum (Lewis et al., 2020). We take a document, replace names and pronouns with either male or female variants and compare summarizer behavior. In the example summaries, entity gender is only explicitly mentioned for the female variant. The model hallucinates that *Melissa Levin* is the *first female executive* of the company.

ficult to disentangle biases that are present in the *summaries* from biases that are already in the *input* documents. We thus propose a procedure that exploits high-quality linguistic annotations to generate mutations of real-world news documents with controllable distribution of demographic attributes.

We make the following contributions:[1]

1. We propose and motivate a number of definitions for bias in text summarization and include novel measures to assess them.

2. We highlight the importance of disentangling *input* driven and *summarizer* driven biases.

3. We conduct a practical gender bias evaluation of both purpose-built summarization models and general purpose chat models for English.

We find that all models score very low on bias in their content selection functions. That is, we find no evidence that the gender of an entity influences the salience of that entity within the summarizers' content models. Where bias occurs, it is often linked to hallucinations. Figure 1 shows a schematic overview of our approach, along with an example of a gender-biased hallucination.

2 Bias in Text Summarization

Bias in NLP is an overloaded term, which is not always used with a clear definition (Blodgett et al., 2020). Before we continue, we thus need to establish our expectations for an unbiased summarizer.

One approach chosen, for example, by Liang et al. (2022) is to require that all demographic groups receive equal representation in the generated summaries, following an *equality of outcome* paradigm (Hardt et al., 2016). While a valid perspective, it requires models to actively *counteract* biases that might be present in the input documents. This is at odds with faithfully representing their content and would thus likely reduce summarizer utility. We instead expect summarizers to be faithful to the input but to not *amplify* their bias. We define three forms of bias under this setting and discuss their harms (Barocas et al., 2017): *inclusion bias*, *hallucination bias* and *representation bias*.

Inclusion bias captures the idea that the (apparent) membership of an entity in some demographic group should not influence how likely that entity is to be mentioned in a summary. If we frame content inclusion in terms of a classification problem over the content units in a document, this corresponds to demanding *equality of opportunity* (Hardt et al., 2016), as opposed to equality of outcome. For example, if both a male- and a female-coded entity are mentioned with otherwise similar salience in a document, the resulting summary should not be more likely to mention the male-coded entity than the female-coded entity, or vice versa. Inclusion bias is thus a property of the summarizer's content selection mechanism. Inclusion bias poses a form of *allocative* harm (Barocas et al., 2017) since it reduces visibility of members of certain groups if, for example, news is consumed through the filter of automatic summarization.

Abstractive summarization systems suffer from *hallucinations* (Kryściński et al., 2019; Cao et al., 2021), that is summary content unsupported by the input. If one demographic group is more likely to feature in them, this would lead to an overrepresentation of this group and entail harms similar to inclusion bias. We call this *hallucination bias*.

The above measures can not capture all kinds of possible bias. As an additional canary, we fi-

[1]Code is available in the supplementary materials.

nally also measure *Representation Bias*, which intuitively measures any kind of summary deviation based on which groups are mentioned in the input. A system exhibits representation bias if it produces different summaries for similar content that relates to different groups. This includes content only included for some groups, entities having different salience in the summary, and differences summary quality. By definition, the presence of any other biases, except hallucination bias, requires the presence of representation bias, but it does not necessarily entail any harms itself. In English texts, for example, we would expect some level of gender representation bias for grammatical reasons.

We want to emphasise that we do not claim that our definitions are universal. They specifically assume we want a summarizer that faithfully reflects the input, regardless of any potential biases therein.

3 Bias Measures

We operationalize our bias measures for a set of demographic groups G. Note that, while in our experiments we only instantiate G as a pair of two groups, all measures generalize to multiple groups.

3.1 Inclusion: Word Lists

A common way to measure bias in text generation is via word lists. For example, Liang et al. (2022) use word lists to evaluate gender bias in LLMs for a variety of tasks, including summarization on CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018). We also assume word lists W_g that identify mentions of each relevant demographic group $g \in G$, and refer to these words as *identifiers* in the remainder of this work.[2]

We then compute the frequency of identifiers in W_g in the set of summaries S: $\text{cnt}(W_g, S)$, deriving an empirical distribution over group identifier frequency $P_{\text{obs}}(g) = \frac{\text{cnt}(W_g,S)}{\sum_{g' \in G} \text{cnt}(W_{g'},S)}, g \in G$. The bias measurement is the total variation distance between P_{obs} and a reference distribution P_{ref}.

As Liang et al. concentrate on equality of outcome (see Section 2), they set P_{ref} as uniform. We instead compute P_{ref} on the source documents to measure *inclusion bias*.

3.2 Inclusion: Entity Inclusion Bias

While word lists are a convenient tool for measuring bias in a general setting, we expect that, in summarization, entities may often be a useful proxy for determining bias. As stated in Section 2, the content selection function of a system without inclusion bias should not be influenced by the group membership of entities[3] in the input. More formally:

$$\begin{aligned} \forall v_i, v_j \in G : p(e \in S | g(e) = v_i, e \in D) \\ = p(e \in S | g(e) = v_j, e \in D) \end{aligned} \tag{1}$$

where $e \in D, e \in S$ indicates that an entity e is mentioned in the source document and summary, respectively and $g(e) = v_i$ indicates that entity e is marked as a member of a demographic group v_i.

We quantify this as the maximum odds ratio between the inclusion probability of two demographic groups. This allows us to compare summarizers with different overall entity density in their summaries. Let $p_{v_i} = p(e \in S | g(e) = v_i, e \in D)$. The inclusion bias score then is

$$\max_{v_i, v_j \in G} \frac{\frac{p_{v_i}}{1-p_{v_i}}}{\frac{p_{v_j}}{1-p_{v_j}}} - 1 \tag{2}$$

where an unbiased system receives a score of 0.

3.3 Hallucination: Entity Hallucination Bias

We operationalize *hallucination bias* by demanding that the probability of a hallucinated entity belonging to a particular demographic group is the same for all groups:

$$\begin{aligned} \forall v_i, v_j \in G : p(g(e) = v_i | e \notin D, e \in S) \\ = p(g(e) = v_j | e \notin D, e \in S) \end{aligned} \tag{3}$$

We measure the total variation distance between $p(g(e) | e \notin D, e \in S)$ and the uniform distribution. We choose the latter since hallucinations introduce new entities, as opposed to reproducing input entities.

3.4 Representation: Distinguishability

Representation bias demands indistinguishability of summaries generated for similar inputs that discuss different demographic groups. We operationalize it by creating a classifier to identify which group is discussed in the input from the summary.

Let S be a set of summaries generated from inputs where both content and mentioned groups are independently distributed. Let $u_i =$

[2]We use the same male and female lists as Liang et al. (see Appendix A) but our formulae are not list-dependent.

[3]We use *entity* exclusively with reference to persons

Corpus	Male	z	Female	z
CNN/DM	league	33.75	ms	51.61
	the	33.75	men/women	39.81
	season	33.64	father/mother	38.52
	club	29.62	’	34.36
	united	29.14	i	33.16
	against	29.07	he/she	32.96
	mr	27.96	baby	32.27
	game	27.76	miss	32.02
	win	27.01	clinton	31.36
	team	25.87	husband	30.49
XSum	mr	28.20	ms	45.49
	(	22.41	men/women	38.63
	)	22.40	mrs	24.40
	shot	16.66	male/female	21.30
	league	16.20	children	19.22
	season	16.12	boys/girls	16.81
	half	16.09	health	15.69
	box	15.70	husband	15.50
	club	15.58	father/mother	14.98
	united	15.18	parents	14.88

Table 1: Ten most male/female associated words in CNN/DM and XSum, with z-scores. Tokens with a slash indicate normalized tokens. For example, *mother/father* is much more frequent in female majority documents.

$\frac{1}{|S_{g(s_i)}|-1}\sum_{s_j \in S_{g(s_i)} \setminus s_i} \text{sim}(s_i, s_j)$ be the average similarity between a summary s_i and all summaries $S_{g(s_i)}$ that have been generated for inputs with the same demographic group that is predominant in s_i. Similarly, let $\bar{u}_i$ be the same for the set of summaries generated for different demographic groups. We say s_i is distinguishable if $u_i > \bar{u}_i$ and compute the distinguishability score as the zero-centered accuracy score of the resulting classifier:

$$\frac{2}{|S|}\sum_{i}^{|S|} \mathbf{1}(u_i > \bar{u}_i) - 1 \quad (4)$$

The metric is parameterized by a similarity function. We use cosine similarity with two representations: A bag of words based representation, and a dense representation derived from Sentence BERT[4] (Reimers and Gurevych, 2019). To avoid distinguishability via simple grammatical cues and names, we replace all pronouns with a gender neutral variant (*they/them* etc.) names with the markers `FIRST_NAME`/ `LAST_NAME`.

4 Input Documents are Already Biased

All proposed measures, except hallucination bias, require us to isolate the effect of a particular demographic group in the input. However, with real world data it is difficult to disentangle *input* driven biases from biases introduced by the *summarizer*.

To demonstrate, we investigate the frequency of gender identifiers from our inclusion score word lists W_g on CNN and XSum *inputs*. We find that 62%/74% of identifiers are male for CNN/XSum, i.e. men are mentioned at a much higher rate.

While this simple frequency issue would be mitigated by our base-rate sensitive formulation of inclusion bias, we find that the underlying issue goes beyond just mention frequency. To demonstrate this, we split the data into two sets, one, where the frequency of female identifiers is higher, and one, where the frequency of male identifiers is higher. We then apply the Fightin' words method (Monroe et al., 2017) with an uninformative Dirichlet prior ($\alpha = 0.01$) to identify words that have a significantly different frequency between male and female texts.[5] Since the word lists are part of our classifier, we replace each pair of male/female words with a special marker.[6] Results in Table 1.

Ignoring the titles (*Mr./Mrs./Ms.*), we see that a number of words have highly significant z-scores ($z \gg 1.96$). Specifically, in both corpora the male documents are much more likely to mention sports related words[7], while documents with more female identifiers have much higher occurrence of words related to family like *husband, children* etc.

We demonstrate the consequences of biased input by examining word inclusion bias of clearly biased summarizers. We consider two content-agnostic baselines: **Random** selects three random sentences. **Lead** selects the first three. We also study two content-aware summarizers. For this we classify every article as either mentioning more *family* or more *sport* based keywords or neither (*unknown*). Classification details can be found in Appendix B. **Topic** randomly samples one, three or six sentences when the article is classified as family, unknown, or sport respectively. **Sexist** selects three sentences to maximize the frequency of male identifiers for sport and of female identifiers for family articles, acting randomly otherwise. The latter is clearly the most biased, while neither **Random** nor **Lead** can, by construction, *amplify* bias. Any bias in **Topic** is a correlation of topics with gender in the input, not due to the algorithm.

We evaluate with word list inclusion bias, since

[4]We use the all-MiniLM-L6-v2 model.

[5]This corresponds to computing the log-odds ratio of token frequencies with a small smoothing factor and then dividing by their standard deviation to receive a z-score.

[6]We ignore the pronouns *him/her/his/hers* in this context due to the POS ambiguity of *“her”*.

[7]This includes the parentheses, which are frequently used in sport reporting, e.g. for results.

	CNN/DM		XSum	
	# Docs	%F	# Docs	%F
Total Docs	11,490	34%	11,334	26%
# Sport	4,222	14%	3,712	14%
# Family	4,317	49%	2,330	36%
Alg.	Unf.	Adj.	Unf.	Adj.
Random	0.15	0.02	0.24	0.00
Lead	0.12	0.00	0.23	0.00
Topic	0.26	0.14	0.29	0.05
Sexist	0.02	0.10	0.20	0.04

Table 2: *First half:* Num. of documents and % of female identifiers per topic. *Second half:* word list inclusion scores of our simulation experiment. *Unf.* and *Adj.* indicate uniform and adjusted reference distribution.

we neither have reliable entity annotation for the CNN/DM or XSum corpora, nor, as our analysis shows, an independent distribution of content and gender as required for distinguishability. Results in Table 2 highlight that: a) Without base rate correction, **Random**, **Lead** and **Topic** appear highly biased, while **Sexist** the least biased. The latter is a consequence of it barely decreasing female representation in sport documents, where representation is already low in the input, but boosting it in summaries for family related articles. b) Even with base rate correction, **Topic** scores higher on bias than **Sexist**, which clearly does not represent the bias of the underlying algorithms.

5 Experimental Setup

5.1 Dataset

We identify three options for creating inputs that avoid the issues outlined in the previous section: 1. Subsampling of existing datasets, 2. Generation of artificial datasets using a LLM, as in Brown and Shokri (2023), 3. Rule-based transformations.

We reject 1, since it requires us to know beforehand which biases exist. Similarly, we avoid LLM data, since it is well known that it is subject to biases itself (Liang et al., 2022). We thus decide on a rule-based approach using linguistic annotations.

Given a corpus C with named entity and coreference information, we create input documents by replacing first names, pronominal mentions and titles of gendered entities to make them read as male or female. Following Parrish et al. (2022), we use popular first names in the 1990 US census (United States Census Bureau, 1990). We leave last names the same to minimize modifications[8]. This allows us to create realistic inputs with controlled gender distribution (see example in Figure 1). We refer to documents from C as *original*, whereas we refer to the modified documents as *inputs*.

We create two variants of the corpus: For C_{loc}, we locally balance gender within each input by assigning half of all entities as male and the other half as female. We use it for inclusion and hallucination bias, since it allows competition between genders for inclusion/hallucination. For C_{glob}, we assign each entity in an input the same gender and instead balance the number of purely male vs. female inputs. We use it for representation bias, since it makes it easy to identify which content is caused by which entity gender assignments. We compute distinguishability within the summaries generated from inputs derived from the same original.

We use the newswire portion of OntoNotes[9] (Weischedel, Ralph et al., 2013) as C so we can avoid the use of coreference resolution that might itself be biased (Rudinger et al., 2018). For both C_{loc} and C_{glob}, we generate 20 inputs for each of the 683 documents in OntoNotes with at least one gendered entity, resulting in 13,660 inputs. Additional details in Appendix D.

5.2 Entity Alignment

To compute entity inclusion and hallucination bias, we require rudimentary cross document coreference resolution between each summary s and input d. OntoNotes gives us access to gold entities E_d and coreference chains in d, but we lack the same annotations in s. We thus first identify all named entities E_s in the summary (without coreference) with a NER tool[10]. While cross-document coreference is difficult (Singh et al., 2011), we exploit the clear correspondence between summary and document in a heuristic instead: We identify the last name for each chain $e_d \in E_d$ by selecting the token that is most frequently in the last position in mentions of e_d (see Appendix D for a detailed description). We align a summary entity e_s to an input entity e_d if e_s contains the last name of e_d. Additionally, we require that any other token in e_s is the first name assigned to e_d during dataset construction or a title.[11] Manual verification of this procedure finds it performs well (Appendix E).

[8] We investigate the effect of this choice in Appendix C.

[9] OntoNotes can be requested from `https://catalog.ldc.upenn.edu/LDC2013T19`.

[10] We use `spacy.io` (Montani et al., 2023)

[11] To avoid incorrectly identifying hallucinations, we additionally require that at least one of the tokens in the entity does not appear in the source to count as hallucinated.

5.3 Identifying Hallucination Gender

While we can identify entity gender of entities that appear in the input by construction, this is not true for hallucinated entities. To compute hallucination bias, we thus design a classification scheme. Since we expect hallucinated entities to often be well known, we first search for a Wikipedia article with a title that exactly matches the entity. If we find one, we determine entity gender by counting gendered pronouns. Otherwise, we fall back to using US census data. We give full detail in Appendix F.

6 Summarizers

We study both **purpose-built** summarizers and **chat** models. For **purpose-built models** we use BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2020), both transformer models fine-tuned on summarization data. We use pretrained XSum and CNN/DM[12] models. For **chat models** we choose Llama-2 chat (Touvron et al., 2023) models with the standard system prompt and sampling procedure. Due to resource constraints, we use the 7b and 13b models. For each summary, we randomly select one prompt from a list of ten prompts designed to elicit summarizing behavior (see Appendix G). We report statistics in Appendix H.

7 Results

Table 3 shows overall scores across all models. We find that all models score low on both inclusion bias measures, indicating that the content selection of all studied models does not carry any significant gender bias *in this particular setting*.

Remarkably, we find that all models carry a bias towards male entities in their hallucinations. We study this in more detail in Section 9.1.

All models show some degree of distinguishablity, with BART summaries showing the most pronounced differences between summaries for male and female coded documents. As noted in Section 2, this is not in itself sufficient to establish whether this leads to harm to any particular group. We analyse this further in Section 9.2.

8 Validating our Measures

To improve the reliability of our results, we additionally validate our methods along three angles: 1) We check whether our modified input documents lead to degraded summary quality 2) We validate that our choice to exclude content words has no mayor impacts on results 3) We test whether our method is capable of detecting entity inclusion and word list bias in clearly biased summarizers.

8.1 Summary Quality

Degradation in summary quality between original and modified articles might be indicative of our inputs being insufficiently natural. We test this using GPT-3.5 with the RTS prompt of Shen et al. (2023) as a reference free metric since it has been shown to be good at identifying low quality summaries. We focus on relevance since we are interested in content selection effects. We find no meaningful quality degradation between summaries on C_{loc} / C_{glob} and originals. Details in App. I, Table 14.

8.2 Content Words

We do not modify gender-specific content words such as *chairman*, which might reduce input naturalness. To check whether this affects results, we manually annotated 100 articles with how content words should be altered depending on entity gender and reran experiments. We find that this has no significant effect on observed bias measures. Details in Appendix J.

8.3 Induced Bias Detection

Since we find no inclusion bias in our summarization models, we test whether our method is capable of detecting inclusion bias of clearly biased systems. Specifically, we append "*Please put a particular focus on the women mentioned in the text*" to the Llama-2 13b prompt to induce it to generate biased summaries.[13] Results in Table 4 show that we can clearly detect the induced inclusion bias.

9 Analysis

9.1 Investigating Hallucination Bias

We investigate what kind of entities are hallucinated. Table 5 contains the ten most frequent hallucinations of each model.

There are two types of frequent hallucinations: For the first type, models often insert entities that are related to the time of the original articles, sometimes by "hallucinating" the original name for an entity in spite of the input, or by inserting the first name for entities that are mentioned without first name in the source. The male bias here can thus

[12]Taken from https://huggingface.co

[13]We manually verify that Llama-2 does not refuse this instruction.

	BART		Pegasus		Llama-2 chat	
	CNN	XSum	CNN	XSum	7b	13b
Word List Inclusion	0.00 s: 0.00,0.01 d: 0.00,0.03	0.04 s: 0.02,0.05 d: 0.00,0.11	0.02 s: 0.01,0.03 d: 0.00,0.06	0.04 s: 0.01,0.06 d: 0.00,0.11	0.04 s: 0.02,0.05 d: 0.01,0.06	0.06 s: 0.05,0.07 d: 0.04,0.08
Entity Inclusion	0.02 s: 0.01,0.03 d: 0.00,0.04	0.02 s: 0.00,0.07 d: 0.00,0.11	0.03 s: 0.01,0.04 d: 0.01,0.05	0.01 s: 0.00,0.05 d: 0.00,0.09	0.00 s: 0.00,0.03 d: 0.00,0.03	0.04 s: 0.03,0.06 d: 0.02,0.06
Entity Hallucination	0.39 s: 0.36,0.42 d: 0.28,0.47	0.37 s: 0.37,0.38 d: 0.31,0.42	0.38 s: 0.35,0.40 d: 0.12,0.49	0.31 s: 0.30,0.33 d: 0.22,0.40	0.38 s: 0.35,0.41 d: 0.30,0.45	0.44 s: 0.42,0.46 d: 0.40,0.47
Distinguishability (Count)	0.21 d: 0.19,0.24	0.24 d: 0.21,0.26	0.16 d: 0.13,0.18	0.15 d: 0.12,0.17	0.05 d: 0.03,0.07	0.08 d: 0.06,0.10
Distinguishability (Dense)	0.22 d: 0.19,0.24	0.24 d: 0.22,0.27	0.16 d: 0.13,0.18	0.15 d: 0.12,0.17	0.04 d: 0.02,0.06	0.08 d: 0.06,0.10

Table 3: Results of our bias measures. In all cases lower scores indicate less evidence of bias. We indicate the 95% bootstrap confidence intervals when resampling original documents (d) and when resampling among the different entity assignments sampled during dataset construction (s). We do not compute (s) for distinguishability, since we can not independently resample scores for input documents generated from the same original document here.

Measure	Llama-2 chat 13b
Word List	0.42 s: 0.41,0.42 d: 0.40,0.44
Entity Inclusion	0.71 s: 0.68,0.74 d: 0.63,0.80

Table 4: Inclusion bias scores on Llama-2 13b prompted to induce an inclusion bias towards female entities.

be attributed to the male-dominant nature of news at article publication times. We rerun our experiments for the C_{loc} case with changed last names to see whether this would alter our conclusions. We find that this has only a limited effect on the hallucination bias. We report detailed results in Appendix C. Our observations link with recent research on *knowledge conflicts* (Wang et al., 2023; Xie et al., 2023), where models may fail to properly reflect answer uncertainty introduced by conflicting evidence in prompt and parametric knowledge. For Llama-2 we manually verify that most hallucinations can be explained in this way.

However, for the purpose-built models, we find a second type of hallucinations that refer to contributors from CNN or the BBC. These usually appear when the summary attributes the text to an author. This is more problematic than historic entities, since they always incorrectly attribute authorship to already potentially well known (mostly male) figures. We find many of these follow repeated patterns. For example, in many instances, BART and Pegasus XSum would generate "In our series of letters from African - American journalists, writer and columnist [name] ...", followed by a description of the article content.

BART	CNN/DM	#	XSum	#
	greene$_u$	91	farai sevenzo$_m$	352
	bob greene$_m$	69	george w. bush$_m$	315
	david frum$_m$	53	mikhail gorbachev$_m$	104
	frum$_u$	47	james baker$_m$	66
	peter bergen$_m$	41	boris yeltsin$_m$	60
	bergen$_u$	41	daniel ortega$_m$	56
	saatchesi$_u$	25	obama$_u$	49
	bynoes$_u$	20	helmut kohl$_m$	40
	frida ghitis$_f$	15	francois mitterrand$_m$	40
	hainis$_u$	12	george h. w. bush$_m$	25
	♯ male	238	♯ male	1465
	♯ female	29	♯ female	212
Pegasus	CNN/DM	#	XSum	#
	frum$_u$	76	boris yeltsin$_m$	60
	david frum$_m$	75	obama$_u$	48
	zelizer$_u$	40	farai sevenzo$_m$	44
	greene$_u$	28	francois mitterrand$_m$	40
	bob greene$_m$	25	richard cohen$_m$	32
	julian zelizer$_m$	20	sharmila tagore$_f$	31
	frida ghitis$_f$	19	helmut kohl$_m$	30
	ghitis$_u$	19	alain juppe$_m$	30
	david weinberger$_m$	8	george w. bush$_m$	25
	bergen$_u$	8	k.$_u$	20
	♯ male	170	♯ male	662
	♯ female	24	♯ female	153
Llama-2 chat	7b	#	13b	#
	mikhail gorbachev$_m$	36	erich honecker$_m$	74
	richard nixon$_m$	29	mikhail gorbachev$_m$	53
	boris yeltsin$_m$	23	richard nixon$_m$	32
	erich honecker$_m$	20	manuel noriega$_m$	32
	mclaren$_u$	20	george h.w. bush$_m$	29
	daniel ortega$_m$	17	daniel ortega$_m$	29
	james baker$_m$	14	walter sisulu$_m$	20
	helmut kohl$_m$	14	mahatma gandhi$_m$	18
	eduard shevardnadze$_m$	12	nelson mandela$_m$	17
	pat nixon$_f$	12	james baker$_m$	17
	♯ male	290	♯ male	545
	♯ female	32	♯ female	35

Table 5: Ten most frequent PERSON named entities without source match in generated summaries. *m*/*f*/*u* indicate entities tagged as *male*/*female*/*unknown* by our name gender classifier (see Sec. 5.3 and App. F)

9.2 Investigating Distinguishability

Distinguishability scores in Table 3 indicate some systematic difference between summaries generated for male and female coded documents, even though we account for expected grammatical differences (see Section 3.4). One possible explanation for this is a difference in summary *quality* between genders which we investigate using reference-free automatic evaluation as in Section 8.1. We report average evaluation scores comparing male and female summaries in C_{glob} in Table 6, finding no quality differences.

System	Avg. Male	Avg. Female	|Diff|
BART XSum	4.30 s: 4.20, 4.41 d: 4.20, 4.41	4.37 s: 4.27, 4.46 d: 4.27, 4.46	0.07 s: 0.01, 0.14 d: 0.01, 0.14
BART CNN/DM	4.84 s: 4.80, 4.89 d: 4.80, 4.89	4.84 s: 4.79, 4.88 d: 4.79, 4.88	0.01 s: 0.00, 0.06 d: 0.00, 0.06
Pegasus XSum	4.24 s: 4.13, 4.35 d: 4.13, 4.35	4.24 s: 4.13, 4.35 d: 4.13, 4.35	0.00 s: 0.00, 0.09 d: 0.00, 0.09
Pegasus CNN/DM	4.59 s: 4.52, 4.67 d: 4.52, 4.67	4.59 s: 4.51, 4.67 d: 4.51, 4.67	0.00 s: 0.00, 0.07 d: 0.00, 0.07
LLAMA 7B	3.50 s: 3.36, 3.63 d: 3.36, 3.63	3.50 s: 3.36, 3.64 d: 3.36, 3.64	0.00 s: 0.00, 0.19 d: 0.00, 0.19
LLAMA 13B	4.98 s: 4.96, 4.99 d: 4.96, 4.99	4.99 s: 4.97, 5.00 d: 4.97, 5.00	0.01 s: 0.00, 0.03 d: 0.00, 0.03

Table 6: GPT3.5 RTS relevance on C_{glob} for summaries on male- and female-only inputs, along with score difference. We compute confidence intervals as in Table 3.

Automatic evaluation can itself be biased and summary quality is only one aspect of representation bias. We thus conduct a manual *qualitative* analysis. We rank input articles in C_{glob} by the (dense) distinguishability of summaries generated for male- and female-coded documents and investigate instances with high distinguishability.

For BART XSum, which has the highest distinguishability, we find there is a pattern where its summaries highlight the gender of women in the context of receiving an appointment to a position of power, but does not do the same for men. We find a total of 12 instances of the bigram "first woman" and an additional 11 instances of the bigram "first female" in the summaries generated by BART XSum, but no instances of "first male" and only a single instance of "first man" (see Figure 1). This not only hallucinates information, but also forms an instance of Markedness (Cheng et al., 2023; Waugh, 1982) since summaries highlight the appointment of women to positions of power as abnormal. We find no similar patterns for the remaining systems.

10 Related Work

While bias in LLMs is the subject of intense research (Sun et al., 2019; Dhamala et al., 2021; Cheng et al., 2023; Srivastava et al., 2023), bias in summarization is underexplored. Liang et al. (2022) include only inclusion bias, measured by word lists. They find strong bias in LLMs, but their measure does not respect the base rate in the input documents (see also Section 2). Their use of CNN/DM and XSum, both highly biased, makes it difficult to attribute this to amplification by models. Brown and Shokri (2023) study summarizers gender bias on GPT-2-generated documents (Radford et al., 2019) using word-embeddings. They find an overrepresentation of men in summaries. In comparison, our data construction reduces the risk of false positives due to input biases and our more differentiated measures suggest hallucination as a likely cause. Zhou and Tan (2023) find summarizers treat articles differently when replacing Biden with Trump and vice versa. While their replacement approach is similar to ours, both their subject of study and measures are highly specific to political bias. Bias has also been observed in tweet and opinion summarization, where contributions by minority groups in the input are underrepresented (Shandilya et al., 2018; Dash et al., 2019; Keswani and Celis, 2021; Olabisi et al., 2022; Huang et al., 2023). In contrast to our bias definition, which focuses on differences in treatment of groups mentioned in the input, here bias is a failure to represent the full *distribution* of opinions and/or authorship.

Ladhak et al. (2023) show models tend to hallucinate entity nationality in biographical summaries. This is consistent with our observation that the most problematic behaviours stem from hallucinations.

Our approach for generating input documents is related to replacement based approaches that generate context that ought to elicit equal behavior for perturbations to the input (Zhao et al., 2018; Parrish et al., 2022), although to the best of our knowledge we are the first to apply such modifications for bias in text summarization.

11 Conclusion

We have introduced definitions that allow us to clearly formulate expectations for what constitutes bias in summarization, along with measures that allow us to detect these biases. We have shown that any measure of *summarizer* bias must account for bias in the input and proposed a rule-based method that allows us to create realistic but unbiased data for studying summarizer bias. While we focus on gender bias in this paper, we present preliminary experiments on ethnicity bias in Appendix K that show that our approach for evaluation and data generation is relevant beyond gender.

Our evaluation of six models indicates that their content selection is not strongly affected by gender bias. However, we caution that content selection in news summarization is known to be subject to easy heuristics like the lead "bias" (Jung et al., 2019). Summaries might be more susceptible to biases in more complicated settings. We find significant bias in hallucinations revealing a connection between hallucination and bias that suggests increasing summarizer faithfulness as a mitigation strategy.

Limitations

Our investigation has a number of limitations. In particular we only study single document summarization for news on English documents. While this is by a large margin the most common setting in summarization research, it can not cover all of the possible applications of summarizers. We also focus mostly on studying binary gender bias, with some preliminary experiments on race.

While we use high-quality linguistic annotations in constructing our templates, issues still arise that limit template creation. We identify the following specific failure cases: 1) our name identification heuristics break down in the few cases where entities are referenced only by their first name 2) named entities are sometimes not linked correctly to coreference chains due to the lack of singleton annotations in OntoNotes. Finally, documents are not always completely natural. In cases where documents mention historic events, names in the article might contradict historical facts. This might limit the generalizability of some of our conclusions.

Ethics Statement

The most significant ethical implication of our work is that our observation that there are few biases in content selection might be misconstrued to imply that these models are generally safe to use. This might lead to less awareness for bias in text summarization. We thus ensure to point out that our conclusions are limited to the particular summarizers and the dataset we used. In particular, it is possible that biases might exist in settings with more complex content selection procedures, such as multi-document summarization.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: From allocative to representational harms in machine learning. In *SIGCIS conference paper*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, Red Hook, NY, USA. Curran Associates Inc. Number of pages: 9 Place: Barcelona, Spain.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Hannah Brown and Reza Shokri. 2023. How (Un)Fair is Text Summarization?

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Inspecting the Factuality of Hallucinated Entities in Abstractive Summarization. *arXiv:2109.09784 [cs]*. ArXiv: 2109.09784.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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). Number of pages: 28 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 172 tex.issue_date: November 2019.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 862–872, New York, NY, USA. Association for Computing Machinery. Number of pages: 11 Place: Virtual Event, Canada.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods* in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. ArXiv:2209.12356 [cs].

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3323–3331, Red Hook, NY, USA. Curran Associates Inc. Number of pages: 9 Place: Barcelona, Spain.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press. Number of pages: 9 Place: Montreal, Canada.

Samhita Honnavalli, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez, and William Yang Wang. 2022. Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation. ArXiv:2205.09830 [cs].

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

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Subaspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3324–3335, Hong Kong, China. Association for Computational Linguistics.

Vijay Keswani and L. Elisa Celis. 2021. Dialect Diversity in Text Summarization on Twitter. In *Proceedings of the Web Conference 2021*, WWW '21, pages 3802–3814, New York, NY, USA. Association for Computing Machinery. Number of pages: 13 Place: Ljubljana, Slovenia.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the Factual Consistency of Abstractive Text Summarization. *arXiv e-prints*, page arXiv:1910.12840. _eprint: 1910.12840.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. ArXiv:2211.09110 [cs].

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, Henning Peters, Paul O'Leary McCann, Jim Geovedi, Jim O'Regan, Maxim Samsonov, Daniël De Kok, György Orosz, Marcus Blättermann, Duygu Altinok, Madeesh Kannan, Raphael Mitsch, Søren Lind Kristiansen, Edward, Lj Miranda, Peter Baumgartner, Raphaël Bournhonesque, Richard Hudson, Explosion Bot, Roman, Leander Fiedler, Ryn Daniels, Kadarakos, Wannaphong Phatthiyaphaibun, and

Schero1994. 2023. explosion/spaCy: v3.6.1: Support for Pydantic v2, find-function CLI and more.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807. Association for Computational Linguistics. Event-place: Brussels, Belgium.

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.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics. Event-place: Vancouver, Canada.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of Extractive Text Summarization. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 97–98, Lyon, France. ACM Press.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 793–803, Portland, Oregon, USA. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí

González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs, stat].

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

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

United States Census Bureau. 1990. Frequently Occurring Surnames from Census 1990 - names file.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*, pages 454–463. Association for the Advancement of Artificial Intelligence (AAAI), Palo Alto, CA.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving Knowledge Conflicts in Large Language Models. ArXiv:2310.00935 [cs].

Linda R. Waugh. 1982. Marked and unmarked: A choice between unequals in semiotic structure. 38(3-4):299–318.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann. 2013. OntoNotes Release 5.0. Artwork Size: 2806280 KB Pages: 2806280 KB.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. ArXiv:2305.13300 [cs].

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Karen Zhou and Chenhao Tan. 2023. Characterizing Political Bias in Automatic Summaries: A Case Study of Trump and Biden. ArXiv:2305.02321 [cs].

A HELM Word Lists

Table 7 shows the word lists W_g we import from Liang et al. (2022) for our word list inclusion measure introduced in Section 3.

Female	Male
she	he
daughter	son
hers	his
her	him
mother	father
woman	man
girl	boy
herself	himself
female	male
sister	brother
daughters	sons
mothers	fathers
women	men
girls	boys
femen[14]	males
sisters	brothers
aunt	uncle
aunts	uncles
niece	nephew
nieces	nephews

Table 7: Male and female word lists reproduced from HELM (Liang et al., 2022).

B Topic Assignment Heuristic

For our demonstration of the effect of input bias in Section 4, we require a transparent way to assign a topic to an input document. Following the observations on gender/topic association in Table 1, we manually select a small number of tokens that we identify as sport or family related. A text is classified by counting the number of occurences for each word list and selecting the majority class. A tie is classified as *unknown*. We list tokens for both categories in Table 8. This allows us to create a deterministic, easy to verify topic assignment. Note that this assignment is purposefully artificial und non-general. It is not intended as a realistic topic classifier, but as a tool to demonstrate how summarizers *might* behave and how this influences bias scores.

C Replacing Last Names

We show entity hallucination scores, along with the other two scores that can be computed on C_{loc} , in Table 9. Results are comparable with the setting that leaves last name intact, with the exception of Llama-2 chat 13b which shows a notable decrease

[14]This is likely a mistake in the original word lists. We reproduce it here for better comparability.

Sport	Family
league	family
season	husband
club	wife
game	father
win	mother
team	children
shot	boys
	girls
	baby

Table 8: Words used for topic identification.

in hallucination score. However, even in the latter case it remains significantly non-zero.

D Corpus Construction

The OntoNotes newswire portion consists of documents from the Wall Street Journal and the Xinhua news agency. We initially consider all documents in the newswire portion for which coreference and named entity annotations are available. From each document, we derive a template which we can then fill with reassigned names and genders in three steps:

1. Identify all coreference chains which refer to a PERSON named entity
2. Determine the first and last name of the entity
3. Identify which mentions of the entity require modifications

In the first step, we consider all coreference chains in the document. If there is any mention that contains a named entity with tag PERSON as a substring, we consider this chain as a candidate for replacement. If the same named entity is part of multiple chains (e.g. because of nested mentions), we link the named entity to the deepest mention that is tagged as IDENT.

Given a chain and with at least one linked PERSON named entity, we try to determine the first and last name of the mentioned entity using a heuristic approach, since there are no annotations for first and last name. We take advantage of two heuristics: 1. titles like Mr./Mrs. are usually followed by a last name 2. mentions with multiple tokens usually contain the first name, followed by the last name

Thus, if a token is preceded by *Mr.*, *Mrs.* or *Ms.* and there is only one other token in the named entity mention, we immediately consider this token as the last name.

Otherwise, we count every token that is the last token in a named entity mention as a possible last name candidate and every token before the last as a possible first name token. Finally, we select the most frequent candidates for first and last name.

In the last step, we consider all mentions of the entity and categorize it into one of the following classes:

Full Name Any mention that contains both first and last name as determined in the previous step

First Name Any mention that contains only the first name

Last Name Any mention that contains only the last name

Pronoun Any mention that is tagged as a PRP or PRP$

Title Any mention that contains a title. We consider *Mr.*, *Mrs.*, *Ms.*, *Sir* and *Lady*.

OntoNotes does not annotate singletons. However, singletons are important since they still require gender adaption to avoid biasing the input. We solve this by treating every PERSON named entity that is not assigned to a chain in the first step as a singleton.

We only consider documents for generation where we find at least one entity with either a first name, gendered personal pronoun or title mention. During the generation of input documents, pronouns and titles are replaced by their male and female equivalents respectively, while first names are replaced with randomly selected names from the 1000 most common first names from the 1990 census data.

To reduce variance due to name selection in C_{loc}, we create pairs of inputs which use the same list of names for both genders but invert the gender assignment of each entity.

E Validation of Alignment Algorithm

To validate that the alignment algorithm in Section 5.2 works as intended, we conduct a manual annotation study. We annotate ten samples each for all systems on both C_{loc} and C_{glob}. This results in a total of 120 input-summary pairs. Since we are

	BART		Pegasus		Llama-2 chat	
	CNN	XSum	CNN	XSum	7b	13b
Word List	0.01 s: 0.00,0.01 d: 0.00,0.04	0.02 s: 0.00,0.04 d: 0.00,0.08	0.03 s: 0.02,0.04 d: 0.00,0.06	0.02 s: 0.00,0.04 d: 0.00,0.09	0.06 s: 0.04,0.07 d: 0.03,0.08	0.07 s: 0.06,0.08 d: 0.05,0.09
Entity Inclusion	0.01 s: 0.00,0.02 d: 0.00,0.03	0.05 s: 0.00,0.09 d: 0.00,0.12	0.01 s: 0.00,0.03 d: 0.00,0.03	0.04 s: 0.00,0.08 d: 0.00,0.10	0.03 s: 0.00,0.06 d: 0.00,0.06	0.02 s: 0.00,0.03 d: 0.00,0.04
Entity Hallucination	0.44 s: 0.41,0.47 d: 0.37,0.48	0.29 s: 0.27,0.31 d: 0.23,0.34	0.41 s: 0.39,0.43 d: 0.23,0.49	0.27 s: 0.25,0.29 d: 0.21,0.33	0.38 s: 0.33,0.43 d: 0.27,0.44	0.32 s: 0.26,0.38 d: 0.18,0.42

Table 9: Results for entity measures computed on C_{loc} with last names altered. We do not report distinguishability, since it requires a corpus in C_{glob} format. We find results are comparable with results without last name alternation. Only Llama-2 13b shows a notable decrease in hallucination score, although it still exhibits strong hallucination bias.

# Input entities	571
# Summary entities	240
# Input entities with alignment in summary	152
# Incorrect entity alignments	**2**
# Summary entities tagged as hallucinated	39
. . . of these with gender classification	17
# Erroneously tagged hallucinations	13
. . . of these with gender classification	**1**

Table 10: Results of our manual annotation of entity alignments. Note that, since we do not have coreference information in the summary, a single input entity can be aligned with multiple summary entities. This may happen case the name is repeated more than once.

interested in validating the alignment, as opposed to the named entity recognizer, we only sample from among all instances where the summary has at least one named entity.

We then manually check the automatic alignment and annotate for each instance:

1. The number of entities in the source that are incorrectly aligned with an entity in the summary.
2. The number of entities in the summary that are erroneously tagged as hallucinated when they are supported by the input. Since hallucinated entities only affect the hallucination bias score when our gender name classification algorithm assigns an apparent gender to the entity, we report how many of these incorrectly tagged entities receive a gender classification and thus might affect the hallucination score. We conduct this annotation on hallucinations before our additional safe-guard requiring at least one token in the entity to not be present in the source.

Results in Table 10 show that our alignment procedure generally works very well. The low number of incorrect alignments can be attributed to the strict matching criteria between summary and source entities. While a third of hallucinations are incorrect, we find that this has little impact on bias scores, since all except one of these hallucinations do not receive a gender classification and thus do not affect the hallucination bias score.

A qualitative analysis reveals that these incorrectly tagged hallucinations are often caused by more complicated coreference settings. For example, five of the incorrectly identified hallucinations are a result of an article discussing a family *"The Beebes"*, which does not get correctly identified as an entity in the input by our approach, since we focus on mentions of individuals. We also find a failure case where the replacement in the input is incomplete, since names are part of nested entities that are not of PERSON type. For example, *"Bush"* in *"The Bush administration"* does not receive a PERSON tag and thus the entity "Bush" can not be aligned to the input. Since in our case these entities are a) not gendered and b) appear in the source document and are thus not taken into account for hallucination bias, this shortcoming of the alignment heuristic does also not affect bias scores.

F Name Classification in Summaries

To determine entity gender in hallucinations, we rely on two separate lookup-based approaches. First, we try to find an English Wikipedia page with a title that exactly matches the named entity detected in the summary (including redirects). To limit false hits, we only consider pages that are in a category that contains the words ”births”, ”deaths” or ”people”. The latter allows matching categories such as ”people from X”, while the first two allow matching categories like ”Y deaths”, where Y is a date. We ignore pages with only a single word in

Male	Female
he	she
him	her
his	hers
himself	herself

Table 11: Pronouns used for entity gender classification.

Please summarize the following old text
Please summarize the following old article
Summarize the following old text
Summarize the following old article
Give a summary of the following old text
Give a summary of the following old article
Give me a summary of the following old article
Give me a summary of the following old text
I need a summary of the following old article
I need a summary of the following old text

Table 12: Prompts used for the Llama-2 Models

the title due to the high likelihood of misidentification.

To determine entity gender for hallucination bias, we use the number of occurrences of the pronouns shown in Table 11 and select the gender with the more frequent pronouns. If we have a tie in the number of pronouns, or if we get conflicting gender predictions due to multiple people with different genders (according to pronoun count) sharing the same name, we classify the gender as unknown. There is a risk that the better coverage of male entities in Wikipedia (Wagner et al., 2015) might influence our bias measure. We thus manually inspect the failure cases of this step and find no evidence that this influences results.

If we do not find a matching entity in Wikipedia, we turn to the 1990 US census first names also used in dataset construction. The census contains gender frequency for each included name. We eliminate duplicates, resolving them to the most frequent gender, if the frequency is at least twice that of the less frequent gender, and eliminating them as ambiguous otherwise. We classify an entity as male, if any token is present in the list of male first names, and as female, if any token is present in the female list. Similarly, we do not classify an entity as either gender if it contains names from both lists.

G Prompts for Llama-2

Table 12 contains the ten prompts we used to elicit summarization behaviour from the Llama-2 models. We specify that the texts/articles are "old" since we found in preliminary experiments that this reduces instances where Llama-2 chat 7b would refuse to summarize articles that contained dates or can be implicitly dated.

H Summary Statistics

Table 13 gives the average number of tokens and entities per summary, as well as the percentage of entities tagged as hallucinated for the summarizers. We find that different summarizers produce summaries of varying length, with XSum summaries being by far the shortest and Llama-2 summaries being the longest. Hallucinations are most frequent on XSum, which is a common observation, since XSum contains hallucinations in gold summaries (Maynez et al., 2020).

I Summary Quality

We are interested in summary quality from two perspectives: a) Do our modifications of the original documents lead to a reduction in summary quality compared to unmodified documents? This might indicate our inputs are unnatural and thus our findings might not generalize; b) Is the distinguishability observed in Table 3 caused by a difference in summary quality for inputs that feature either male or female entities?

Since we do not have access to gold summaries, we use an unsupervised evaluation method. Following the recent success of using large language models in reference-free evaluation for text generation (Liu et al., 2023; Chiang and Lee, 2023; Shen et al., 2023), we use GPT 3.5 to elicit rating for the generated summaries. We prompt the model using the reason-then-score prompt of Shen et al. (2023):[15]

> Score the following Summary given the corresponding Article with respect to relevance from one to five, where one indicates "irrelevance", and five indicates "perfect relevance". Note that relevance measures the Summary's selection of important content from the Article, whether

[15]We use the `gpt-3.5-turbo-1106` model. This model is more recent than the one used in the evaluation of Shen et al., but allows us to fit the entirety of the documents and summaries into the available tokens.

Corpus	C_{loc} Avg. Tok.	C_{loc} Avg. Ent.	C_{loc} % Hal.	C_{glob} Avg. Tok.	C_{glob} Avg. Ent.	C_{glob} % Hal.	Orig Avg. Tok.	Orig Avg. Ent.
BART CNN/DM	60.76 d: 60.17,61.39	0.97 d: 0.88,1.07	4.65 d: 3.06,6.67	60.88 d: 60.29,61.45	0.99 d: 0.90,1.09	4.01 d: 2.46,5.76	60.60 d: 59.90,61.26	1.00 d: 0.90,1.10
BART XSum	23.55 d: 23.10,24.02	0.27 d: 0.23,0.32	51.28 d: 43.10,58.30	23.59 d: 23.09,24.07	0.28 d: 0.24,0.32	47.67 d: 40.18,54.89	22.81 d: 22.31,23.27	0.25 d: 0.21,0.29
Pegasus CNN/DM	56.23 d: 55.10,57.41	0.87 d: 0.79,0.96	3.29 d: 1.60,5.29	56.19 d: 55.07,57.36	0.86 d: 0.78,0.95	3.32 d: 1.56,5.42	55.29 d: 53.98,56.72	0.79 d: 0.71,0.87
Pegasus XSum	24.69 d: 23.75,25.77	0.22 d: 0.18,0.25	33.69 d: 25.67,40.95	24.74 d: 23.71,25.78	0.22 d: 0.19,0.26	32.09 d: 24.90,39.31	22.90 d: 22.13,23.66	0.19 d: 0.16,0.23
LLama2 7b	164.40 d: 162.70,166.07	0.97 d: 0.88,1.06	2.97 d: 1.97,4.06	165.38 d: 163.60,167.07	0.99 d: 0.89,1.08	3.18 d: 2.16,4.43	175.52 d: 172.75,178.11	1.22 d: 1.08,1.37
LLama2 13b	163.80 d: 161.48,166.01	1.55 d: 1.40,1.69	2.95 d: 1.93,4.18	163.87 d: 161.35,166.10	1.56 d: 1.41,1.71	2.89 d: 1.91,3.90	166.86 d: 163.84,169.93	1.63 d: 1.47,1.79

Table 13: Average number of tokens and entities, and percentage of all entities tagged as hallucinated. We compute bootstrap confidence intervals for all values on the document level.

> the Summary grasps the main message of the Article without being overwhelmed by unnecessary or less significant details.
>
> Article: {article}
>
> Summary: {summary}
>
> Provide your reason in one sentence, then give a final score:

For each system, we evaluate all 683 summaries generated from the original documents which are used as templates for C_{loc} and C_{glob} . For C_{loc} and C_{glob} themselves we conserve resources and only evaluate summaries generated for two randomly selected inputs, resulting in 1366 ratings per system.

We report the average score for each summarizer in Table 14 to validate the impact of our modifications on overall summary quality, whereas we discuss the impact of input document gender on summary quality in the main text in Section 9.2, Table 6.

We find that, while there is a small reduction in score for 4 out of 6 systems, performance is very similar between original and modified documents, with the latter score falling within less than one standard deviation of the original score. This indicates that our modification of the input documents does not lead to meaningful degradation in summary quality.

J Content Words

J.1 Motivation

Our automatic template generation procedure only changes names and pronominal mentions, leaving content words unchanged. This can lead to unnatural occurrences, such as *Chairman Diane Sasser*, when *Chairwoman Diane Sasser* would be more appropriate. To check whether this is an issue in our experiments, we manually extend the automatically derived templates to also modify content words.

J.2 Annotation Procedure

Since we found in preliminary experiments that many articles do not require any manual intervention, we first run an automatic filter over our dataset to identify candidate articles for annotation. We use an extended variant of our word list W_g of Liang et al. (2022) reproduced in Table 15. We then randomly sampled from these articles until we found 100 instances where at least one text span required manual intervention to adapt to entity gender.

During annotation, we identified text spans which should change in accordance with the gender of an entity in the document and which words should be used in either case (e.g. generating *chairman* or *chairwoman* depending on the gender of the entity occupying that position). We also considered the case where multiple entities might influence the realization of a particular word, like *brothers*. In these cases, we also specify a neutral variant (e.g. siblings) to be used in case the referenced entities have different genders.

J.3 Results

We report results in Table 16. We find scores for modified inputs are very close to original scores when taking into account confidence intervals and exhibit the same trends. However, we note that the relatively small number of inputs makes confidence intervals relatively wide.

K Ethnicity Bias

To demonstrate our method is useful beyond studying binary gender bias, we conduct additional experiments investigating ethnicity bias in summarization and its interaction with gender. We focus on biases related to black/white associated names in an US-American context, since this is well known to be subject to biases in language models (Cheng et al., 2023; Parrish et al., 2022; Liang et al., 2022).

We construct five datasets to study this bias: One, where gender is randomly assigned, and all four

System	C_{loc}	Std.	C_{glob}	Std.	Original	Std.
Pegasus XSum	4.23	1.45	4.24	1.45	4.28	1.42
Pegasus CNN/DM	4.57	1.03	4.59	1.01	4.70	0.89
BART XSum	4.32	1.38	4.34	1.37	4.30	1.40
BART CNN/DM	4.81	0.66	4.84	0.60	4.86	0.60
Llama-2 chat 7B	3.50	1.83	3.50	1.82	3.85	1.68
Llama-2 chat 13B	4.99	0.18	4.98	0.22	4.99	0.15

Table 14: GPT-3.5 RTS scores for summaries generated on C_{loc} , C_{glob} and on original documents. For C_{loc} , C_{glob} we evaluate summaries for two inputs each for each article ($n = 1366$). For the original documents, we evaluate all summaries ($n = 683$). We find only minor differences in quality between summaries on C_{loc} / C_{glob} and original documents, indicating that our procedure does not result in systematic degradation of summary quality.

Female	**Male**
daughter	son
mother	father
woman	man
girl	boy
female	male
sister	brother
daughters	sons
mothers	fathers
women	men
girls	boys
females	males
sisters	brothers
aunt	uncle
aunts	uncles
niece	nephew
nieces	nephews
wife	husband
wives	husbands
actress	actor
actresses	actors
chairwoman	chairman
chairwomen	chairmen
mum	dad
mums	dads
waitress	waiter
waitresses	waiters
mistress	lover

Table 15: Extended word list used to identify candidate articles for annotation.

possible combinations of entity race and binary gender. We use the black and white name dictionary of Parrish et al. (2022) for constructing entity names. We change both first and last names, since both are relevant in communicating entity ethnicity. Due to the small name inventory (10 per race and gender), we can not generate instances for all documents. We thus only consider originals where we can generate a full set of 20 inputs for all settings, leaving us with 12,240 instances per dataset.

Since word lists for ethnicity bias typically rely on last names, we do not compute word list bias and only compute inclusion bias. We also opt not to compute hallucination bias, since we want to avoid constructing a classifier that attempts to identify race of entities mentioned in the summary.

Table 17 shows that most models exhibit no entity inclusion bias, with the exception of BART XSum, which exhibits inclusion bias across all investigated settings. In all cases BART XSum prefers to include black associated names in the summary. We find a pattern in distinguishability that is similar to our observations in gender bias, where the purpose-built models all exhibit relatively high distinguishability, whereas the Llama-2 models receive much lower distinguishability scores. Interestingly, we find that for all models that have significantly non-zero distinguishability, it is highest when black and white coded entities are assigned opposite genders. Similarly, for BART XSum, inclusion bias is highest in these settings, although we note that none of the differences are significant.

L Computational Infrastructure

We ran most inference on a single node using four RX6800 GPUs. For Llama-2 13b we ran some additional experiments on a single A100 GPU.

	BART		Pegasus		Llama-2 chat	
	CNN	XSum	CNN	XSum	7b	13b
Word List	0.00 s: 0.00,0.02 d: 0.00,0.06	0.00 s: 0.00,0.04 d: 0.00,0.15	0.02 s: 0.00,0.04 d: 0.00,0.08	0.06 s: 0.03,0.11 d: 0.00,0.21	0.08 s: 0.05,0.11 d: 0.03,0.14	0.04 s: 0.03,0.06 d: 0.01,0.09
Word List (Orig.)	0.04 s: 0.02,0.05 d: 0.00,0.11	0.06 s: 0.02,0.09 d: 0.00,0.21	0.05 s: 0.03,0.07 d: 0.00,0.13	0.09 s: 0.06,0.12 d: 0.00,0.25	0.05 s: 0.02,0.08 d: 0.01,0.11	0.05 s: 0.04,0.07 d: 0.02,0.10
Entity Inclusion	0.03 s: 0.00,0.07 d: 0.00,0.07	0.05 s: 0.00,0.16 d: 0.00,0.22	0.04 s: 0.01,0.08 d: 0.00,0.09	0.01 s: 0.00,0.10 d: 0.00,0.24	0.01 s: 0.00,0.08 d: 0.00,0.09	0.06 s: 0.02,0.09 d: 0.01,0.10
Entity Inclusion (Orig.)	0.01 s: 0.00,0.05 d: 0.00,0.08	0.16 s: 0.05,0.29 d: 0.02,0.39	0.02 s: 0.00,0.06 d: 0.00,0.07	0.02 s: 0.00,0.10 d: 0.00,0.19	0.01 s: 0.00,0.08 d: 0.00,0.07	0.02 s: 0.00,0.06 d: 0.00,0.07
Entity Hallucination	0.30 s: 0.19,0.41 d: 0.00,0.48	0.32 s: 0.29,0.34 d: 0.11,0.46	0.45 s: 0.00,0.46 d: 0.00,0.48	0.36 s: 0.32,0.39 d: 0.00,0.48	0.35 s: 0.22,0.46 d: 0.20,0.45	0.46 s: 0.00,0.48 d: 0.00,0.49
Entity Hallucination (Orig.)	0.26 s: 0.16,0.36 d: 0.01,0.44	0.33 s: 0.31,0.35 d: 0.15,0.47	0.00 s: 0.00,0.00 d: 0.00,0.00	0.33 s: 0.30,0.36 d: 0.00,0.49	0.17 s: 0.02,0.32 d: 0.01,0.39	0.41 s: 0.33,0.47 d: 0.12,0.48
Distinguishability (Count)	0.42 d: 0.35,0.49	0.42 d: 0.36,0.50	0.28 d: 0.21,0.35	0.24 d: 0.16,0.31	0.05 d: 0.00,0.11	0.18 d: 0.09,0.25
Distinguishability (Count) (Orig.)	0.27 d: 0.21,0.34	0.33 d: 0.27,0.41	0.19 d: 0.13,0.24	0.20 d: 0.13,0.27	0.11 d: 0.06,0.16	0.12 d: 0.07,0.17
Distinguishability (Dense)	0.40 d: 0.33,0.47	0.41 d: 0.34,0.48	0.27 d: 0.20,0.34	0.23 d: 0.16,0.31	0.03 d: -0.03,0.09	0.16 d: 0.08,0.24
Distinguishability (Dense) (Orig.)	0.27 d: 0.21,0.34	0.33 d: 0.25,0.41	0.19 d: 0.13,0.24	0.22 d: 0.15,0.29	0.05 d: 0.00,0.09	0.07 d: 0.01,0.12

Table 16: Results on our manually extended variants of C_{loc} and C_{glob} . Since our annotations cover only a relatively small subset of the whole corpus, we also report the scores of summaries generated for the same inputs without content word modification for comparison (Orig.). We find that almost all scores fall within their respective confidence intervals.

	BART		Pegasus		Llama-2 chat	
Gender Assignment	CNN	XSum	CNN	XSum	7b	13b
Entity Inclusion Bias						
Random	0.01 s: 0.00,0.03 d: 0.00,0.04	0.17 s: 0.11,0.24 d: 0.08,0.29	0.04 s: 0.00,0.09 d: 0.00,0.11	0.02 s: 0.00,0.04 d: 0.00,0.05	0.01 s: 0.00,0.04 d: 0.00,0.04	0.01 s: 0.00,0.03 d: 0.00,0.03
Black Male/White Female	0.03 s: 0.01,0.05 d: 0.00,0.05	0.19 s: 0.13,0.26 d: 0.10,0.32	0.04 s: 0.00,0.09 d: 0.00,0.12	0.02 s: 0.00,0.03 d: 0.00,0.05	0.05 s: 0.02,0.08 d: 0.01,0.08	0.02 s: 0.00,0.04 d: 0.00,0.04
Black Male/White Male	0.05 s: 0.03,0.07 d: 0.02,0.08	0.11 s: 0.06,0.18 d: 0.03,0.22	0.08 s: 0.03,0.14 d: 0.01,0.16	0.05 s: 0.03,0.07 d: 0.03,0.08	0.02 s: 0.00,0.05 d: 0.00,0.06	0.01 s: 0.00,0.03 d: 0.00,0.03
Black Female/White Male	0.03 s: 0.01,0.05 d: 0.00,0.06	0.24 s: 0.17,0.30 d: 0.13,0.36	0.05 s: 0.00,0.10 d: 0.00,0.15	0.01 s: 0.00,0.03 d: 0.00,0.04	0.01 s: 0.00,0.04 d: 0.00,0.04	0.01 s: 0.00,0.03 d: 0.00,0.03
Black Female/White Female	0.01 s: 0.00,0.03 d: 0.00,0.04	0.12 s: 0.06,0.18 d: 0.02,0.22	0.02 s: 0.00,0.07 d: 0.00,0.10	0.04 s: 0.02,0.06 d: 0.01,0.07	0.01 s: 0.00,0.04 d: 0.00,0.04	0.01 s: 0.00,0.03 d: 0.00,0.03
Distinguishability (Count)						
Random	0.19 d: 0.16,0.21	0.23 d: 0.20,0.25	0.16 d: 0.14,0.19	0.10 d: 0.08,0.12	0.01 d: -0.01,0.03	0.04 d: 0.02,0.06
Black Male/White Female	0.24 d: 0.22,0.27	0.28 d: 0.25,0.30	0.18 d: 0.16,0.21	0.13 d: 0.11,0.16	0.03 d: 0.01,0.05	0.04 d: 0.01,0.06
Black Male/White Male	0.20 d: 0.17,0.22	0.25 d: 0.22,0.27	0.15 d: 0.13,0.17	0.09 d: 0.06,0.11	0.02 d: -0.00,0.04	0.02 d: 0.00,0.04
Black Female/White Male	0.23 d: 0.21,0.26	0.30 d: 0.27,0.33	0.26 d: 0.24,0.29	0.19 d: 0.16,0.21	0.03 d: 0.01,0.05	0.05 d: 0.03,0.07
Black Female/White Female	0.21 d: 0.18,0.23	0.24 d: 0.22,0.27	0.22 d: 0.19,0.24	0.12 d: 0.09,0.14	0.01 d: -0.01,0.03	0.01 d: -0.01,0.04
Distinguishability (Dense)						
Random	0.16 d: 0.13,0.19	0.21 d: 0.19,0.24	0.17 d: 0.15,0.19	0.10 d: 0.08,0.13	0.02 d: -0.00,0.04	0.03 d: 0.01,0.05
Black Male/White Female	0.24 d: 0.22,0.27	0.28 d: 0.26,0.31	0.19 d: 0.17,0.22	0.13 d: 0.10,0.15	0.01 d: -0.01,0.03	0.03 d: 0.01,0.05
Black Male/White Male	0.19 d: 0.17,0.22	0.23 d: 0.21,0.26	0.16 d: 0.13,0.18	0.09 d: 0.07,0.12	0.02 d: 0.00,0.04	0.02 d: 0.00,0.04
Black Female/White Male	0.24 d: 0.22,0.27	0.29 d: 0.27,0.32	0.26 d: 0.24,0.29	0.18 d: 0.16,0.20	0.04 d: 0.02,0.06	0.05 d: 0.03,0.08
Black Female/White Female	0.21 d: 0.19,0.24	0.23 d: 0.21,0.26	0.22 d: 0.19,0.24	0.12 d: 0.10,0.15	0.02 d: -0.00,0.04	0.02 d: 0.00,0.04

Table 17: Bias scores for entities with black/white associated names with different gender assignments. *Random* assigns gender uniformly at random, independently of race. We find that only BART XSum has a slight inclusion bias towards entities with black-associated names. In distinguishability, we find that purpose-built summarizers exhibit some degree of distinguishability, whereas Llama-2 chat models score low along this axis as well.