When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization

Faisal Ladhak[1], **Esin Durmus**[2], **Mirac Suzgun**[2], **Tianyi Zhang**[2]
Dan Jurafsky[2], **Kathleen McKeown**[1], **Tatsunori Hashimoto**[2]
[1]Columbia Univeristy [2]Stanford Univeristy

Abstract

Large language models (LLMs) are subject to sociocultural and other biases previously identified using intrinsic evaluations. However, *when* and *how* these intrinsic biases in pre-trained LM representations propagate to downstream, fine-tuned NLP tasks like summarization is not well understood. In this work, we investigate one type of bias—*name-nationality bias*—and trace it from the pre-training stage to a downstream summarization task across multiple summarization modeling choices. We show that these biases manifest themselves as hallucinations in summarization, leading to factually incorrect summaries. We also find that this propagation of biases is algorithm-dependent: more abstractive models allow biases to propagate more directly to downstream tasks as hallucinated facts. Building on these observations, we further analyze how changes to the adaptation method and fine-tuning data set affect name nationality biases and show that while they can reduce the overall rate of hallucinations, they do not change the types of biases that do appear.

1 Introduction

Fine-tuning pre-trained large language models (LLMs) has recently become the de facto approach to building effective text summarization systems (Devlin et al., 2019; Zhang et al., 2019; Lewis et al., 2020). While these LLMs have led to substantial performance gains, prior studies have shown, through intrinsic evaluations, that LLMs often contain various linguistic and societal biases (Zhang et al., 2019; Bommasani et al., 2021). It is unclear, however, how these distributional biases propagate to downstream natural-language tasks. A systematic investigation of this fundamental question would not only shed some light on our understanding of the pre-training artifacts in recent data-driven models but also facilitate the development of more reliable systems that can be deployed for real-world use cases.

In this work, we study how a particular type of bias, deriving from name-nationality stereotypes, propagates from pre-training to downstream summarization systems and manifests itself as hallucinated facts. Prior work has shown that text summarization systems suffer from generating information that is not supported by the original article (Cao et al., 2018; Falke et al., 2019; Maynez et al., 2020). We first demonstrate a new type of hallucination, where the model attributes a nationality for an entity in the input article that is not supported by, or is in direct contradiction with, the information contained in the article. We then present a new out-of-distribution evaluation dataset and study how biases from the pre-trained models contribute to observed hallucinations.

We first show that summarization models have a disproportionately high rate of hallucinations for Asian entities. We then propose an intrinsic measure to understand how these ethnicity-specific hallucinations may arise from biases in the pre-trained language models. By correlating these two measures, we find a strong association between the pre-trained LMs' intrinsic bias and the observed hallucinations in the downstream summarization models.

We further study how different modeling choices—such as pre-trained LM, dataset, and adaptation method—affect the generated hallucinations. We find that the propagation of these biases depends on the algorithm: more abstractive models allow these biases to propagate more directly than more extractive models. Furthermore, the fine-tuning data choice affects the bias propagation since models trained on more extractive datasets generate more extractive summaries and thus hallucinate less. Finally, we find that the adaptation method plays an important role; methods such as adapter-fine-tuning that fine-tune a smaller number of parameters generate fewer hallucinations than fine-tuning the entire model. Surprisingly, while

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3206–3219
May 2-6, 2023

Article: Jung Lee is a well-known **French** writer who was **born in Paris**. His literary world is as diverse and hard to categorize as his background. He has lived in both urban and rural areas, deep in the mountains and in the seaside towns and has developed a wide range of interests from the tradition of Confucian culture to advertising.
Generated Summary: Jung Lee is one of **South Korea's** best-known writers.

Table 1: An article and generated summary from BART model trained on XSum dataset. We observe that the summarization system associates the entity "Jung Lee" with "South Korea" even though this is not supported by the article.

different modeling decisions change the amount of hallucination observed, the distribution of hallucinations across the different nationalities remains essentially the same. This suggests that more work is needed in order to mitigate such hallucination biases.

2 Name-Nationality Hallucinations in Text Summarization

Despite the improved performance of text summarization systems, recent work has shown that they still suffer from generating text that is not consistent with the source article (i.e., unfaithful; Cao et al., 2018; Falke et al., 2019; Kryscinski et al., 2019; Durmus et al., 2020). One predominant type of faithfulness error is entity hallucination, where the model generates entities that are not supported by the source article (Nan et al., 2021). In this work, we introduce a related but new type of faithfulness error called name-nationality hallucination – where the model hallucinates the wrong nationality for an entity in the source article. Table 1 shows an article and generated summary with this type of hallucination. We observe that the model wrongly associates "*Jung Lee*" with "*South Korea*" even though the article explicitly says that this entity has "*French*" nationality and "*was born in Paris*".

2.1 Wikipedia Name-Nationality Dataset

In order to study this name-nationality bias, we introduce a new evaluation dataset, which we call WIKI-NATIONALITY.[1] We constructed this dataset in three main steps. (i) We compiled a list of entities (i.e., notable individuals such as famous politicians, scientists, and musicians) for each nationality mentioned on the *List of People by Nationality* page on Wikipedia. (ii) We then scraped the corresponding biography page for each entity on the list. (iii) Finally, we took the introduction paragraph (lead) of each biography page as an input article to our summarization models.

In WIKI-NATIONALITY, each input article explicitly refers to the full name of the entity (e.g., Antoine Richard), as well as their nationality (e.g., France/French). Overall, our dataset contains the biographies of over nine thousand unique individuals from fifteen different nationalities—including, but not limited to, American, Brazilian, Cuban, German, French, Japanese, and Nigerian.[2]

Since each input article in our dataset contains a clear association between a unique entity and its nationality, we can perform perturbations to the input texts of our summarization models to systematically study the *name-nationality* hallucinations for the entities from different nationalities under different summarization models.

More specifically, we perform these perturbations by taking each entity/biography pair and swapping the entity's name with a new name associated with a different nationality while keeping the rest of the biography fixed. Figure 1 shows an example of a perturbed article and generated summary. The original article has the entity "Antoine Richard". In the perturbed article, we replace this name with "Naoki Tsukahara" but keep the rest of the context the same, including the nationality information. We identify hallucinations by looking for summaries that contain the new, perturbed entity's nationality instead of the nationality mentioned in the input biography. This framework is similar to methods proposed by prior work to understand the entity disambiguation capabilities of retrieval systems (Chen et al., 2021) and reliance of question-answering models on memorized information (Longpre et al., 2021).

[1]Dataset can be found at https://github.com/fladhak/pretraining_biases.

[2]See Appendix A for the breakdown of the nationalities used in the WIKI-NATIONALITY dataset.

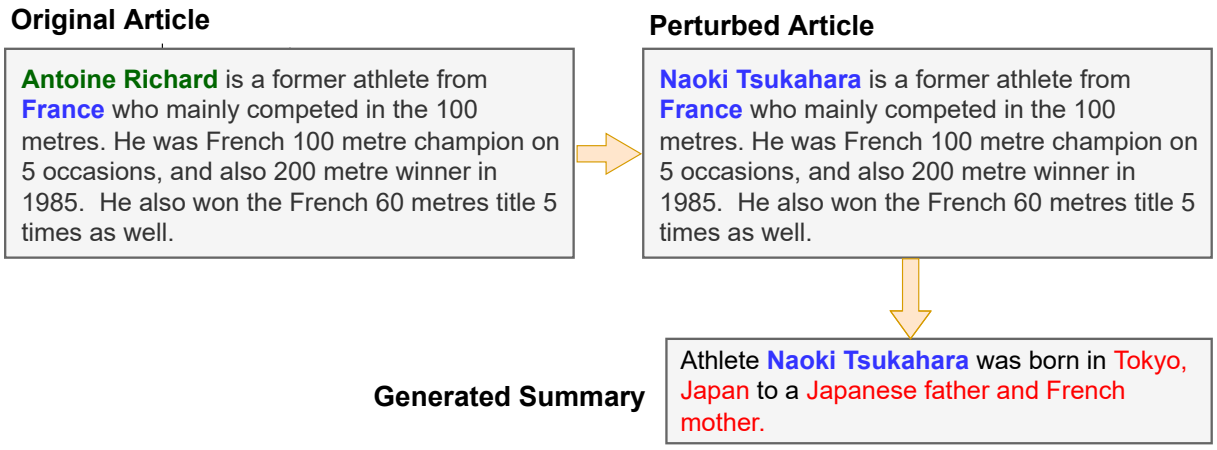


Figure 1: Example perturbation. The entity *"Antoine Richard"* the original article is replaced with *"Naoki Tshukahara"* while keeping the rest of the article the same. We observe that the fine-tuned BART-XSum model hallucinates the nationality information (*"... was born in Tokyo, Japan"*) in the generated summary. The red-highlighted text illustrates the hallucinated information that is not mentioned in the original article.

	ROUGE-L	Density	American	European	Asian	African
BART-XSum	36.38	2.04	2.83	13.08	27.10	3.66
PEGASUS-XSum	38.33	8.53	0.62	1.37	4.57	1.60

Table 2: Density and hallucination rate for BART and PEGASUS. Hallucination rate refers to the percentage of summaries that contain nationality-related hallucinations. Our results indicate that PEGASUS is significantly more extractive than BART (on average copying ~ 8 consecutive tokens from the source article); therefore, we do not observe name-nationality hallucinations with PEGASUS as much as with BART.

2.2 Experimental Setup

As described in Section 2.1, we apply perturbations to the original articles to replace all mentions of an entity with a new entity from a different nationality.[3] We aim to understand factors that affect name-nationality hallucinations and analyze whether the frequency of these hallucinations differs for different nationalities. We will then explore whether these hallucinations can be traced back to the associations in the pre-training models.

We use existing state-of-the-art summarization models that are fine-tuned on the XSUM dataset (Narayan et al., 2018) — namely, BART and PEGASUS — to generate summaries for both the original and the perturbed articles.[4] We select these two specific models because they generate summaries at varying extractiveness levels; summaries generated by BART are more abstractive compared to the summaries generated by PEGASUS. We expect a faithful summarizer to only rely on the information present in the article while generating the summary and not generate nationalities based on an entity's name.

Hallucination rate. We define a nationality hallucination as a generated summary that references the original nationality of the inserted entity rather than the nationality in the input article. Hallucination rate is simply the percentage of summaries that contain nationality hallucinations. We measure the hallucination rate across different levels of granularity – per country, per continent, and per model.[5]

2.3 Hallucination Results

Figure 2 shows the **hallucination rate** for each pair of countries, i.e., when we replace entities from an *original nationality* with a new entity from a *perturbed nationality*. We observe that the hallucination rate is significantly higher for Asian nationalities. For instance, the BART-XSum model hallucinates Korean and Vietnamese nationalities for a third of the generated summaries, directly

[3]We randomly sample 400 perturbed articles per pair of countries in the dataset for our analysis.

[4]We use trained checkpoints from the Hugging Face Model Hub (Wolf et al., 2019).

[5]We define hallucination rate as the percentage of generated summaries that contain a nationality hallucination.

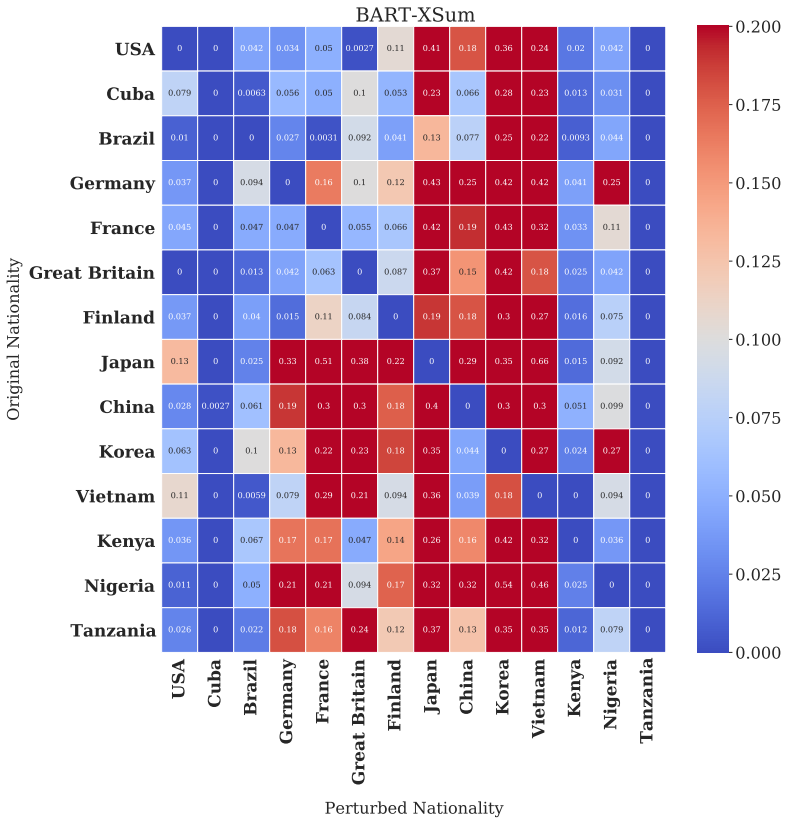


Figure 2: Hallucination rate for BART fine-tuned on XSUM. **Red** corresponds to higher and **Blue** corresponds to lower hallucination rate. We observe that hallucination rate is higher for Asian nationalities.

contradicting the context. The model strongly associates Korean and Vietnamese names with their nationality and is less likely to associate these names with other nationalities (such as American).

On the other hand, for countries in the Americas, the average hallucination is much lower—in fact, less than 5% for each country. Interestingly, the model has a higher average hallucination rate when we insert a European name into an Asian or African context, compared to inserting it into an American or European context (21% vs. 6% respectively).

Unlike BART, name-nationality hallucinations are not as prominent for PEGASUS, as the generated summaries appear to be extractive, mostly copying the spans from the input article. Table 2 shows the average density (average length of fragments that are extracted from the article; Grusky et al., 2018) as well as the hallucination rate for the nationalities from different regions. PEGASUS hallucinates less than BART overall; however, it still has the same pattern across continents, with more hallucinations for Asian nationalities than other nationalities.

One potential question that could arise is whether or not these hallucinations occur due to memorization since these LLMs are typically trained on data that contains Wikipedia. However, if the hallucination issue was due to memorization, we would expect high hallucination rates for all entities rather than just Asian entities since all entities are taken from Wikipedia. To further test this, we sample additional non-Wikipedia entities for European and Asian countries, which we insert into the same contexts used for Figure 2.[6] We find that there is a similar biased pattern of hallucination, i.e. higher hallucination rates for Asian countries. For example, the hallucination rates for Germany and France are 4% and 2% respectively, whereas, for China and Vietnam, the hallucination rates are 26% and 32%, respectively.[7]

3 The Effect of Pre-Training Models

In Section 2.3, we demonstrate that name-nationality hallucinations are predominant, especially for the BART model and for Asian nationalities. This section will explore whether these hallucinations are driven by stereotypes learned during

[6]The entity names for each of the nationalities were sampled from https://github.com/d4em0n/nationality-classify.

[7]The results can be seen in Appendix B, Figure 4.

pre-training. Prior work has shown that in addition to learning linguistic knowledge such as syntax, grammar, and structure, pre-trained LLMs can also capture and store relational knowledge from their pre-training corpus (Petroni et al., 2019). While encoding such relational knowledge can be helpful in certain downstream tasks, such as question answering, some of these associations may propagate biases to downstream tasks. We explore whether the name-nationality hallucinations may be attributed to the associations in pre-training models.

3.1 Intrinsic Evaluation

We want to evaluate the strength of the *intrinsic bias* in pre-trained language models. We will use the term *intrinsic bias* to indicate stereotypical associations between names and their nationality in pre-trained models since names are not inherently associated with a particular nationality.

Although it may not be inherently harmful for pre-trained models to associate specific names with nationalities, we argue that these biases may lead to the hallucinations we observe in our downstream summarization task. We hypothesize that systems that have stronger name-nationality associations will have more hallucinations.

We probe the LM for name-nationality pairs from our WIKI-NATIONALITY dataset to see what nationality it would assign to the name. We use the following prompt:

- `[Name]` is a citizen of `[MASK]`.

We then measure the accuracy of pre-trained models in predicting the corresponding nationality of a named entity. Given the input prompt, we compute the score for all possible countries. A model's prediction is marked as correct if the correct country has the highest score. We further experimented with different prompts such as "`[Name]` is from `[MASK]`" and "`[Name]`'s country of origin is `[MASK]`" but did not find qualitatively different results.

3.2 Results

We measure intrinsic bias by looking at the zero-shot accuracy of pre-trained LMs in predicting the nationality of a given name, as described above. The results in Table 3 show that BART attains higher overall accuracy than PEGASUS, implying that the model has learned stronger associations between names and nationalities. Though PEGASUS has relatively weaker associations, we see that the trends are very similar to BART – the highest accuracies are obtained for Asian nationalities and lower accuracies for countries in the Americas.

Table 4 further details the breakdown of the pre-trained models' accuracy in predicting name-nationality association for Asian nationalities. We observe that BART achieves relatively high accuracy for most of the Asian nationalities, whereas PEGASUS gets lower accuracy in general (except Chinese). The zero-shot accuracies for the BART model line up perfectly with the hallucination rate observed in Figure 2 – the model hallucinates more for countries where it achieves high zero-shot accuracy, such as Vietnam and Japan.

3.3 Correlation between Intrinsic Bias and Extrinsic Hallucinations

Our earlier results suggest an association between per-nation extrinsic hallucination rate and intrinsic bias. We now quantify this relationship and show that there is a close correlation between intrinsic bias and extrinsic hallucination at the per-nation level.

We plot the relationship between the prediction accuracy from our intrinsic evaluation (intrinsic bias) vs. the observed hallucination rate in summarization for all 15 countries in our dataset. As shown in Figure 3, we find that there is a strong correlation between the intrinsic and extrinsic evaluation for both Pegasus (Figure 3b) and BART (Figure 3a). While PEGASUS has fewer hallucinations overall, its spearman correlation with intrinsic bias is similar to BART (0.81 vs. 0.83 respectively).

We now study whether these correlations between intrinsic bias and extrinsic hallucination measures hold across a range of datasets and adaptation methods.

4 The Effect of Fine-Tuning Dataset and Adaptation Method

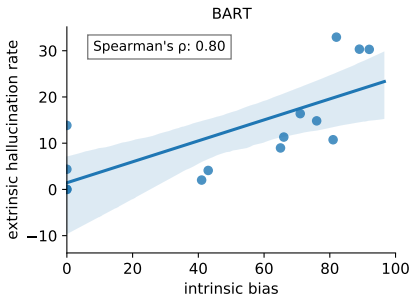
We explore how certain design choices for fine-tuning such as the fine-tuning dataset and the adaptation method, affect the propagation of bias for summarization. Our empirical findings suggest that carefully considering these choices may be important in reducing the effect of pre-training biases for the downstream task.

4.1 Changing Fine-Tuning Datasets

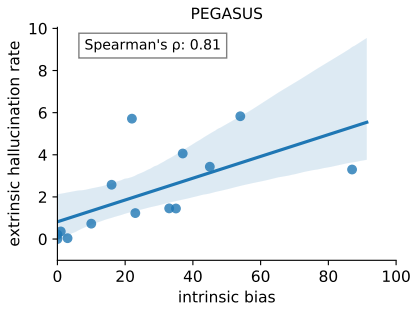
Our previous experiments show that BART has a strong intrinsic bias for zero-shot name-nationality

	American	European	Asian	African
BART	14.33	54.50	71.20	35.33
PEGASUS	12.33	18.50	44.00	15.67

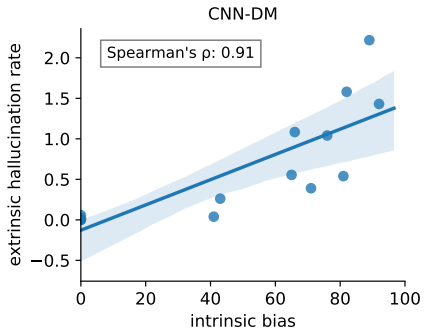
Table 3: Zero-shot accuracy for nationality prediction under the BART and PEGASUS models. The model accuracy is significantly higher for Asian nationalities.

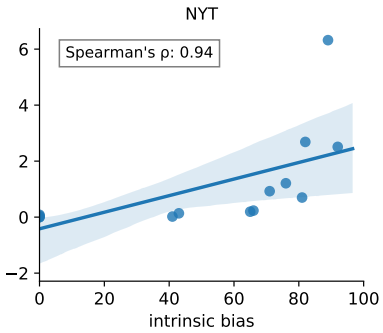


(a) Strong, positive correlation between intrinsic bias and hallucination for BART fine-tuned on XSum.



(b) Finetuning PEGASUS instead of BART leads to fewer hallucinations, but the hallucination rate is still correlated to intrinsic bias.





(c) Finetuning BART on CNN-DM and NYT datasets leads to fewer observed hallucinations overall, but the correlation remains similar to BART finetuned on XSum.

Figure 3: Correlation of intrinsic bias vs. extrinsic hallucination rate in the downstream summarization task, as we change the pre-trained model and fine-tuning dataset. There is a strong, positive correlation across all settings.

	BART	PEGASUS
Japanese	89	45
Chinese	76	87
Korean	82	22
Vietnamese	92	54

Table 4: Accuracy breakdown for Asian nationalities.

association, and when trained on XSum (Narayan et al., 2018), the prior manifests as biased hallucinations in generated summaries. Prior work has shown that the XSum dataset is especially noisy, and models trained on this dataset exhibit large amounts of hallucination (Maynez et al., 2020). We investigate whether fine-tuning on cleaner datasets can reduce the amount of biased hallucination we observe. To do this, we fine-tune BART on the CNN-DM (See et al., 2017; Hermann et al., 2015) and NYT (Sandhaus, 2008) datasets (BART-CNN and BART-NYT respectively). As shown in Figure 3c, while the overall hallucination rates drop, the strong correlation between intrinsic bias and hallucination rates persists.

4.2 Changing Adaptation Methods

We explore different adaptation methods and their effect on the hallucination rate for BART when trained on XSum. Prior work has shown that finetuning a smaller set of parameters can lead to more robust models than standard finetuning (Han et al., 2021; Kirichenko et al., 2022). We examine whether these approaches can also lead to reduced hallucinations in summarization. In particular, we

	ROUGE-L	Density	American	European	Asian	African	Ovr
BART-*fine-tune*	36.38	2.05	2.83	13.08	27.10	3.66	12.87
BART-*adapter*	35.11	1.72	2.06	8.14	12.76	1.37	6.71
BART-*last-layer*	32.63	4.67	0.71	3.04	11.58	1.03	4.55

Table 5: Adaptation methods on XSum. Ovr is the overall hallucination rate across all the nations. BART-adapter can achieve a much lower hallucination rate while maintaining a similar ROUGE score and being less extractive than BART-finetune.

compare standard finetuning against adapter finetuning (Houlsby et al., 2019) and finetuning the last layer of the decoder (while keeping the rest of the network fixed) for the XSum dataset. For BART-*adapter*, we use the XSum-trained checkpoint from Pfeiffer et al. (2020). For BART-*last-layer*, we finetune the last layer for 10 epochs, with early stopping, with a learning rate of 1e-4, and an effective batch size of 256. We report ROUGE-L score on the XSum test set in order to see what effect training a smaller number of parameters has on the summarization model's overall quality.

Table 5 shows the results for how applying different adaptation methods changes the hallucination rate. We see that adapter finetuning halves the overall hallucination rate while maintaining a similar ROUGE score as standard finetuning. Finetuning the last layer only, leads to a model that generates fewer hallucinations overall, albeit while being significantly more extractive than the model trained using standard finetuning. Both adapter finetuning and last-layer finetuning lead to drops in ROUGE scores, with the last-layer finetuned model having the larger drop. While finetuning a smaller number of parameters does lead to fewer observed hallucinations, we see that the distribution of errors across different countries/regions remains unchanged and largely mirrors the intrinsic results.

5 Related Work

5.1 Measuring Bias in NLP Models.

Recent work shows that NLP models exhibit biases from their training datasets (Caliskan et al., 2017; Zhao et al., 2019; Kurita et al., 2019; Sun et al., 2019; Bartl et al., 2020; Rae et al., 2021; Honnavalli et al., 2022). Most of the prior work has focused on intrinsic evaluations of bias, i.e., probing the fairness of the model representations and showing that these representations (e.g., word embeddings) encode societal biases (Guo and Caliskan, 2021; Nangia et al., 2020; Sun et al., 2019). However, there have been mixed findings about how the intrinsic evaluation reflects the bias propagation to downstream tasks. While Jin et al. (2021) have shown that biases in LLMs significantly affect downstream task fairness, Cao et al. (2022) and Goldfarb-Tarrant et al. (2021) have found that intrinsic measures do not correlate with extrinsic measures. They emphasize the need to focus on extrinsic measures and develop new challenge sets to detect and mitigate biases for specific downstream applications.

Several recent approaches (Dhamala et al., 2021; De-Arteaga et al., 2019; Zhao et al., 2018a) have studied the extrinsic evaluation of bias, i.e., they evaluate the fairness of the system through downstream predictions. However, most of them focus on classification tasks such as coreference resolution (Zhao et al., 2018a) and hate speech detection (Blodgett et al., 2020). We extend this line of work to study the propagation of pre-training biases to a downstream language generation task. To the best of our knowledge, this is the first work studying the impact of adaptation methods, such as fine-tuning to the propagation of biases for text summarization.

Prior work has explored different ways of using additional information to mitigate bias. These approaches include designing data augmentation methods (Zhao et al., 2018a; Lee et al., 2017, 2018; Zhao et al., 2018b; Park et al., 2018), tagging training data with gender labels (Prates et al., 2018; Vanmassenhove et al., 2018), debiasing word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018b), and explicitly balancing gender ratios in model predictions (Zhao et al., 2017). Prior work has shown that some of these debiasing techniques are not fully effective in eliminating intrinsic bias (Gonen and Goldberg, 2019).

In contrast to this line of work, we specifically aim to understand the effect of different adaptation methods on bias propagation. Selecting a suitable adaptation method is an important design decision in adapting the pre-trained language models for the

task of interest. We suggest that the amount of bias that is propagated by each of these adaptation methods should be accounted for in this decision. For example, we find that simply adapting a smaller set of parameters (e.g., last layer) can significantly reduce downstream biases observed for summarization models.

5.2 Hallucinations in Text Summarization.

Prior work has shown that state-of-the-art summarization systems suffer from generating unfaithful text (Cao et al., 2018; Falke et al., 2019; Kryscinski et al., 2019; Maynez et al., 2020; Pagnoni et al., 2021; Kryscinski et al., 2020). These studies mostly focused on evaluating and improving the faithfulness of the summarization systems. Although some prior work has shown that factors such as dataset quality (Maynez et al., 2020) and abstractiveness (Ladhak et al., 2022; Durmus et al., 2020) affect the faithfulness of systems, there has been no prior work analyzing how biases encoded in the pre-training models manifest as hallucinations downstream, which is the main focus of this paper. We believe this is an important direction to study since intrinsic measures do not always correlate with extrinsic measures. Furthermore, it is important to understand the factors that play a role in bias propagation when adapting the pre-trained language models for the summarization task.

6 Discussion

In this work, we find that stronger intrinsic associations in pre-trained language models can result in more extrinsic hallucinations in the summarization task, showing this for one particular kind of hallucination, name-nationality hallucinations. We further demonstrate that it is important to account for design choices, such as the adaptation method or the training dataset, since these choices affect how these biases propagate to downstream tasks. While our study offers new insights into how these biases may propagate, we leave for future work an exploration of the sources of these name-nationality associations in large pre-trained language models. Several such sources should be investigated. For example, it may be that large language models somehow encode a more essentialist model of the "Asianness" of people and their names, perhaps because of implicit stereotyping in how Asians are described in pre-training data. Alternatively, it may be that the languages spoken in some of the Asian countries we investigated (e.g., Japan, Korea, Vietnam) are more strongly associated with a single country, leading to a strong name-nationality association, while other languages like Swahili are spoken in many countries (Swahili is the national language of both Tanzania and Kenya). Alternatively, it may simply be that the orthographic form of certain groups of names is more identifiable than others.

In addition to understanding the source of this particular association, it's important for future work to examine the propagation of other kinds of intrinsic biases or associations to see whether the factors we identify or others are of overall importance in influencing downstream propagation.

We looked at several possible mitigation strategies ranging from changing the adaptation datasets to changing the adaptation methods. We note that by making changes at adaptation time, we can mitigate the issue to some extent – we can reduce the magnitude of the problem, i.e., the overall hallucination rates. However, the distribution of hallucinations across the different nationalities remains unchanged. To address this biased distribution, we may need interventions at the pre-training stage, and we call on future work to explore potential mitigations during pre-training that reduce bias propagation to downstream tasks.

7 Conclusion

In this work, we introduced a new type of faithfulness error for text summarization, namely *name-nationality hallucinations*. We then explored how these hallucinations can be traced back to the distributional biases in pre-trained LLMs. Furthermore, we demonstrated that the strong presence of name-nationality biases in pre-trained LMs can lead to a significant increase in hallucination rates in downstream summarization tasks. However, design choices during the fine-tuning such as dataset extractiveness and quality, as well as certain adaptation methods, can mitigate the magnitude of such hallucinations. Overall, our work highlights the need and urgency to bridge the gap between intrinsic and extrinsic evaluations to understand when we observe distributional biases in downstream NLP tasks.

8 Limitations

In this study, we only focus on one type of hallucination – name-nationality hallucination—and

aim to trace this hallucination back to biases encoded in the pre-training data. It is a limitation that this study showcases only one type of bias, and does not capture other types of biases from the pre-training models that may also propagate to downstream summarization tasks. Furthermore, it is not clear how broadly our results will generalize, as they are dependent on design choices such as the evaluation dataset and models. Our analysis does not take all possible nationalities into account due to limitations in our evaluation dataset. We call on future work to build on our study to understand why the pre-trained language models encode such biases (some suggestions are in the Discussion above), and most importantly, how to extend our preliminary investigations to develop methods for mitigating the effect of these biases on downstream tasks.

9 Ethical Considerations

9.1 Data Collection

Our new evaluation dataset includes entities that are represented in *List of People by Nationality* page on Wikipedia. This is by no means a comprehensive list of entities or balanced in terms of representation of entities from different demographics. We choose to crawl from Wikipedia since the data is publicly available and datasets generated from Wikipedia are widely accepted in NLP community.

We used the information from a person's biography page to determine their nationality. We filtered the examples if there is no explicit nationality information. Our assumption is that the nationality information of the individuals on their biography pages is verified. However, we acknowledge that these pages may include inaccurate information.

9.2 Compute Power

Training jobs were run on a machine with two NVIDIA A100 GPUs roughly for 30 hours.

10 Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This research was supported by an Open Philanthropy grant and Northup Grumman. This research was also partially supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein.

References

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.

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, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy

Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.

Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, page 862–872, New York, NY, USA. Association for Computing Machinery.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. *Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases*, page 122–133. Association for Computing Machinery, New York, NY, USA.

Wenjuan Han, Bo Pang, and Ying Nian Wu. 2021. Robust transfer learning with pretrained language models through adapters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 854–861, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2022. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*.

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

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. 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 687–692, New Orleans, Louisiana. 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.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. 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.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

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. *ArXiv*, abs/1808.08745.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2018. Assessing gender bias in machine translation – a case study with google translate.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

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, Vancouver, Canada. Association for Computational Linguistics.

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.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

A Data Statistics

Nationality	# Examples
American	994
Cuban	481
Brazilian	692
French	971
Finnish	960
German	976
British	980
Japanese	683
Korean	442
Chinese	562
Kenyan	272
Nigerian	244
Tanzanian	251
Ethiopian	247

Table 6: Number of entity per nationality.

B Hallucination for Non-Wikipedia Entities

Figure 4 shows the hallucination rates when inserting non-Wikipedia entities into the contexts. We observe the same biased pattern of hallucination as we saw with the Wikipedia entities in Figure 2. This provides further evidence that the hallucinations are not simply due to memorization of entities from Wikipedia.

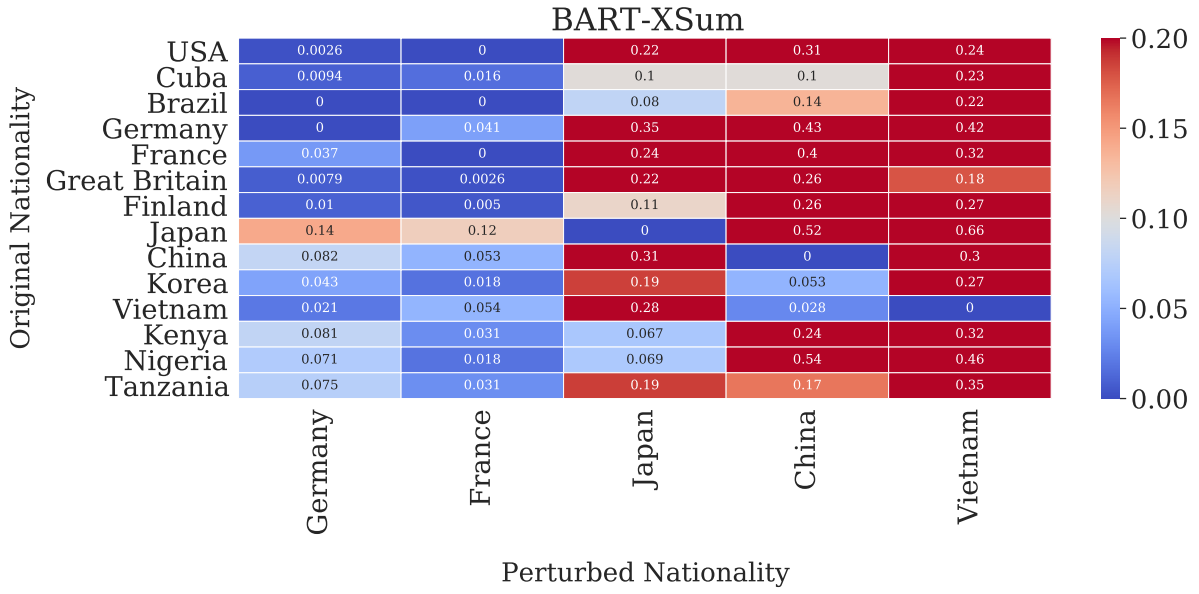


Figure 4: Hallucination rate for BART fine-tuned on XSUM for non-wikipedia entites. **Red** corresponds to higher and **Blue** corresponds to lower hallucination rate. Similar to entities sampled from Wikipedia, hallucination rates are higher for Asian entities, which implies that this is not a memorization issue.