Speciesist Language and Nonhuman Animal Bias in English Masked Language Models

Anonymous ACL submission

Abstract

Warning: This paper contains examples of offensive language, including insulting or objectifying expressions.

Various existing studies have analyzed what social biases are inherited by NLP models. These biases may directly or indirectly harm people, therefore previous studies have focused only on human attributes. If the social biases in NLP models can be indirectly harmful to humans involved, then the models can also indirectly harm nonhuman animals. However, no research on social biases in NLP regarding nonhumans exists. In this paper, we analyze biases to nonhuman animals, i.e. speciesist bias, inherent in English Masked Language Models. We analyze this bias using template-based and corpus-extracted sentences which contain speciesist (or non-speciesist) language, to show that these models tend to associate harmful words with nonhuman animals. Our code for reproducing the experiments will be made available on GitHub.

1 Introduction

Recently, in the field of Natural Language Processing (NLP), Masked Language Models (MLMs) using Transformers (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), widely contributed to the state-of-the-art methods in downstream tasks. However, existing studies suggest that these models inherit social biases (Sun et al., 2019; Blodgett et al., 2020). Such biases cause differences in accuracy between majority and minority attributes (e.g. Romanov et al., 2019) and negative generalizations, e.g. in text generation (Liu et al., 2020; Sheng et al., 2019, 2021; Garimella et al., 2021).

The studies of social bias in NLP target gender (e.g. Bolukbasi et al., 2016; Caliskan et al., 2017), race (e.g. Manzini et al., 2019), religion and ethnicity (e.g. Li et al., 2020) and so on, all of which assume human attributes. However, to the best of the authors' knowledge, there are no similar bias studies on **nonhuman animals**.

In this paper, we use templates, corpus-extracted sentences and pre-trained MLMs to investigate if the bias regarding nonhuman, i.e. **speciesist bias**, is inherent in MLMs trained on English corpora.

The bias we investigate in this paper is the **representational bias**, following the classification of Sun et al. (2019) and Blodgett et al. (2020). Currently, nonhuman animals do not use the NLP system directly, so we do not need to consider the idea of, e.g. "performance against the social group of nonhuman animals". On the other hand, we think that we should respect nonhuman animals **for their own sake, not for the sake of humans** (cf. Owe and Baum, 2021), for the reasons described below, and therefore we should study, for example, insulting associations with nonhuman animals and negative stereotyping against them.

1.1 Ethical Discussion: Nonhumans and NLP

There may be more possible criticisms of the research objectives of this paper. The first criticism is that there is no ethical problem with the existence of harmful bias to nonhuman animals.

However, we should give equal consideration to interests and should not discriminate based on **who** has the interests (Singer, 2015). Even if one does not accept this idea, most people would agree that nonhuman animals deserve some moral consideration (Owe and Baum, 2021). If this is true, then it is important to study the biases that are harmful to nonhuman animals.

The second potential criticism is that even if nonhuman animals deserve some moral consideration, NLP models with speciesist bias do not harm them because they do not use it directly. However, we think it is important to study the speciesist bias of NLP models for three following reasons.

First, if NLP systems with a speciesist bias are popularized in our society, the bias of the NLP system may affect us and thereby indirectly harm animals (in human animal cases, see Bender et al., 2021) [1]. For example, if an NLP system generates speciesist sentences, the speciesist bias may propagate to readers who read the sentences, and they may acquire an implicit discriminatory bias against nonhuman animals. As we discuss in Section 2.2, we are already discriminatory against nonhuman animals, but we think this phenomenon should not be reinforced.

Second, the representational speciesist bias should be considered unwarranted in itself, even if it does no direct harm (Blodgett et al., 2020). The use of language that is insulting to or demeaning nonhuman animals, as described in Section 2.2, is wrong in itself (cf. Hellman, 2008), even if nonhuman animals never recognize the expression.

Third, the biases inherent in word embeddings reflect social biases which exist in our cognition, beliefs and social structures (Caliskan et al., 2017; Garg et al., 2018; Joseph and Morgan, 2020). Therefore, analyzing the speciesist bias in word embeddings and corpora can contribute to research about the influence of this bias on our cognition and society.

For these reasons, we think that it is important to study the speciesist bias in NLP.

2 Related Work

2.1 Social Bias in Language Models

Existing studies (Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019) have shown that social biases are inherent in word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). Moreover, some other studies have found that also Masked Language Models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) inherit social biases. In these studies, social biases of contextualized word embedding have been intrinsically assessed using template sentences (Bartl et al., 2020; Hutchinson et al., 2020; Kurita et al., 2019; May et al., 2019; Tan and Celis, 2019; Webster et al., 2020; Silva et al., 2021), corpus sentences (Basta et al., 2019; Guo and Caliskan, 2020; Zhao et al., 2019) and manually generated paired sentences (Nadeem et al., 2021; Nangia et al., 2020).

[1] "Stochastic parrots" in the title of Bender et al. (2021) is an example of specisist language use.

2.2 Speciesism and Language

Speciesism is "the unjustified comparatively worse consideration or treatment of those who do not belong to a certain species." (Horta and Albersmeier, 2020, p.3). Nonhuman animals, as sentient beings, deserve equal consideration with human animals (Singer, 2015, p.40), and we should not discriminate against nonhuman animals. However, we do so, for example by eating their flesh or conducting experiments on them (Singer, 2015, ch.2, 3).

We also treat nonhuman animals as inferior beings or objects in our language use. For instance, "terming a woman a 'dog'" insults all women indirectly and also insults all dogs directly (Dunayer, 1995, p.12). Usual referring to nonhuman animals as "it" or "something," or using "that" or "which" as relative pronouns to indicate nonhuman animals are examples of treating nonhuman animals as objects (Dunayer, 2001, 2003). Dunayer (2001, ch.9) also states that, in the process of slaughtering, people use words such as "harvest", "package" and "process" to hide cruelty.

In addition to research conducted in Animal Ethics field, there are also studies in Corpus Linguistics that analyzed language use regarding nonhuman animals. Jepson (2008) performed discourse analysis on various texts and spoken conversations showing that the word "slaughter" in human context collocates strongly with negative emotions, but lacks such sentiment when used in the context of nonhuman animals. Franklin (2020) also analyzed the use of "killing" terms, such as "kill" and "slaughter", in "People, Products, Pests and Pets" (PPPP)[2] which is an English corpus that contains texts referring to nonhuman animals extracted from various domains such as food-related websites and news articles (Sealey and Pak, 2018).

Existing studies have reported that stylistic biases are reflected in NLP models (Tan et al., 2020; Hovy et al., 2020). Therefore, since the above-mentioned speciesist language and biases in English may be reflected in MLMs, we investigate a possibility of speciesist bias in English MLMs.

3 Experimental setup

The MLMs used in this paper are $\text{BERT}_{\text{LARGE-cased}}$[3], $\text{RoBERTa}_{\text{LARGE}}$[4],

[2] https://animaldiscourse.wordpress.com/
[3] https://huggingface.co/bert-large-cased
[4] https://huggingface.co/roberta-large

DistilBERT$_{\text{base-cased}}$[5] (Sanh et al., 2019) and ALBERT$_{\text{large-v2}}$[6] (Lan et al., 2020), which are widely used in current NLP. We determine animals we focus on in this paper as follows:

1. We collect animal names from "All Animals A-Z List."[7] We focus on only one-term names.
2. We limited the number of animals for this research by choosing only these which names appear on English Wikipedia[8] more than 20,000 times, resulting in 46 animal names in total.

Our hypothesis is that if MLMs recognize different animals by categorizing them, then similar bias will be found for animals in similar contexts. In this paper, we categorize animals who live in farms to be utilized as flesh marking them in ■, nonhuman companions in ■, and other animals in ■ colors, respectively. In Table 1, we show all animal names under investigation, their corresponding colors, and their frequencies in Wikipedia.

4 Bias Analysis by Speciesist and Non-Speciesist Language

In this section, we explain how we evaluate the speciesist bias inherent in MLMs using (1) template-based and (2) corpus-based approaches. The template-based approach is commonly used in bias analysis of NLP models. However, the template-based approach may limit aspects of biases that can be evaluated, depending on the template (Guo and Caliskan, 2020). Therefore, we conduct bias evaluation also by using raw sentences extracted from a corpus.

4.1 Template-based Experiment

The basic template sentence we utilize is **"[PRONOUN] is a [ANIMAL] [REL-PRONOUN] is [MASK]."**, where [PRONOUN] slot indicates a pronoun, [ANIMAL] is an animal name, and [REL-PRONOUN] stands for a relative pronoun.

We evaluate bias toward [ANIMAL] by observing the change of predicted probability of words at the [MASK] token by replacing [PRONOUN] and [REL-PRONOUN]. We use the following combinations of [PRONOUN] and [REL-PRONOUN]:

Table 1: Animal names used in this research and their frequencies in English Wikipedia. The coloring of animal names was done by the authors: ■ refers to "farm" animals, ■ represents popular nonhuman companions and ■ addresses all remaining species.[9]

Animal name	Frequency	Animal name	Frequency
horse	194,363	deer	43,130
turkey	187,079	seal	42,533
fox	176,569	snake	42,323
human	173,145	persian	39,764
fish	142,508	duck	36,828
dog	127,775	swan	36,556
bird	124,463	sheep	34,433
moth	93,670	chicken	34,231
buffalo	91,392	snail	33,725
robin	89,168	bombay	32,819
cat	83,038	frog	31,922
wolf	78,795	crane	31,328
eagle	78,126	penguin	30,769
bear	69,029	rat	28,851
lion	67,774	monkey	28,144
tiger	60,709	falcon	27,843
beetle	54,887	rabbit	27,039
bat	49,445	beaver	26,421
mouse	48,866	pike	25,392
fly	45,411	pig	25,273
newfoundland	44,353	elephant	24,817
tang	44,245	cow	22,563
butterfly	44,096	molly	21,353

- human-describing sentences (hereinafter referred to as "human sentences")
  - *She* is a [ANIMAL] *who* is [MASK].
  - *He* is a [ANIMAL] *who* is [MASK].
- object-describing sentences (hereinafter referred to as "object sentences")
  - *This* is a [ANIMAL] *which* is [MASK].
  - *That* is a [ANIMAL] *which* is [MASK].
  - *It* is a [ANIMAL] *which* is [MASK].
  - *This* is a [ANIMAL] *that* is [MASK].
  - *That* is a [ANIMAL] *that* is [MASK].
  - *It* is a [ANIMAL] *that* is [MASK].

In human sentences, we use "she", "he", and "who", which generally refer to humans. In object sentences, we use "this", "that", "it", and "which", which are generally used for nonhumans. Since pronouns in object sentences are only in the third person equivalently, only the third person pronouns "she" and "he" are used in human sentences.

Our hypothesis here is that the characteristics of the words that are filled in "[MASK]" will change among animals that are often referred to in the speciesist language and others that are not. For

[5] https://huggingface.co/distilbert-base-cased

[6] https://huggingface.co/albert-large-v2

[7] https://a-z-animals.com/animals/

[8] We use the Wikipedia dataset downloaded on 01/05/2020 from https://huggingface.co/datasets/wikipedia.

[9] Following Crameri et al. (2020), in this paper we use scientific color map (Crameri, 2021) to include people with diverse color vision.

example, not only humans, but also dogs and cats could be referred to by the non-speciesist language, while "farm animals" (e.g. cow and pig) would be addressed by the speciesist language.

4.1.1 Bias Evaluation by Word Probability Differences

We evaluate the bias against animal names using words with a large change rate of average predicted probability between human and object sentences. It is done by averaging predicted probability of the word filled into the [MASK] token in the template sentences. We also investigate the relationship between animals by clustering them using the agreement rate of words with large probability changes. We perform this experiment as follows:

1. Calculating mean probability $p^{w_i}_{mean_o}(\text{name})$ and $p^{w_i}_{mean_h}(\text{name})$ in object and human sentences, respectively, where name is an animal name and w_i is a token in vocabulary V of the MLM (i.e. $w_i \in V$)
2. Calculating how much this probability changes by $\log \frac{p^{w_i}_{mean_o}}{p^{w_i}_{mean_h}}$
3. Ignoring words w_i if (a) both $p^{w_i}_{mean_o}$ and $p^{w_i}_{mean_o} < \frac{1}{|V|}$, or (b) |z-score| of $\log \frac{p^{w_i}_{mean_o}}{p^{w_i}_{mean_h}}$ for each MLM lower threshold[10]
4. Calculating Token-Match-Rate (TMR) among animal names
5. Clustering all animals based on TMR with UPGMA algorithm (Michener and Sokal, 1957).

In step 1, we calculate $p^{w_i}_{mean_{o,h}}$ as follows:

$$p^{w_i}_{mean}(\text{name}) = \frac{1}{|T|}\sum_{s\in T}^{|T|} p(w_i = \text{"[MASK]"}|s(\text{name})) \quad (1)$$

where T is the set of object or human template sentences described above, $s(\text{name})$ is a template sentence filled with an animal name. In step 4, where $S^{(i)}$ and $S^{(j)}$ are the obtained sets of words for the i, j-th animal names after step 3, we calculate $\text{TMR}(i, j)$ between both sets (cf. Webster et al., 2020; Lauscher et al., 2021):

$$\text{TMR}(i, j) = \frac{|S^{(i)} \cap S^{(j)}|}{\min(|S^{(i)}|, |S^{(j)}|)} \quad (2)$$

In step 5, we cluster animal names by using $1 - \text{TMR}(i, j)$ as distance between i, j-th names.

[10]In these experiments we ignore words with |z-score| lower than 1.96. We set this point experimentally in order to obtain significant words.

4.1.2 Bias Evaluation by Sentiment Analysis

In this experiment, we use VADER (Hutto and Gilbert, 2014) for evaluating the sentiment of all words which we obtain from the experiment described in Section 4.1.1. This approach does not take into account context when evaluating sentiment of the words, but we decided to analyze the sentiment of the words themselves, considering the possibility of (non-)speciesist bias in the animal names.

Our hypothesis is that when animals are regarded as objects, they are treated negatively, and therefore more negative words will appear under MASKs in object sentences.

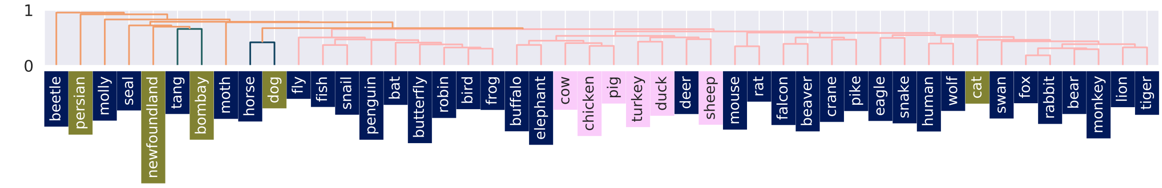
4.2 Corpus-based Experiment

In this section, we explain how the bias is measured in the corpus-based evaluation method. The corpus used in this paper is Books3 (Presser, 2020, see also (Gao et al., 2020)) which totals about 100GB of text and is built only from published books. Thus, it is unlikely to overlap with BookCorpus (Zhu et al., 2015), which contains unpublished books used for the pre-training of MLMs.

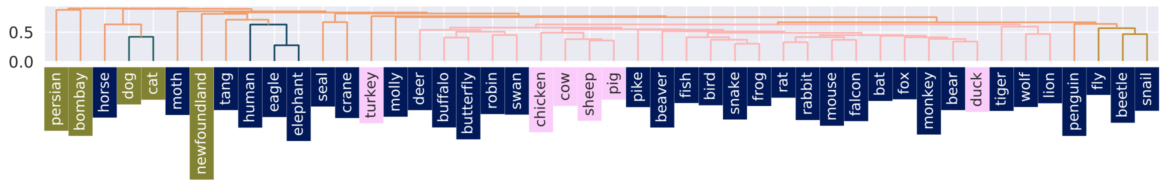
To experiment with corpus-based method, we extract object and human sentences from a given corpus. For the purpose of this research, we extract all corpus sentences that contain relative pronouns referring to animals. We use five relative pronouns: "that", "which", "who", "whose" and "whom". Our assumption is that these relative pronouns can be used to determine whether (non)human animals are treated as objects or humans in the given sentence.

CoreNLP (Manning et al., 2014) is used to extract sentences containing relative pronouns which refer to an animal name. If the speciesist bias exists in CoreNLP, then there may be a difference in referring precision between human and object sentences. Therefore, we asked a native speaker of English to check whether relative pronouns are correctly referred to an animal name in ten sentences (for each pronoun) randomly extracted from Book3. As a result, one sentence containing "who", and two with "whom" have been marked as incorrect, and all remaining 47 sentences have been judged as having correct references. It suggests that the precision of the parser for this task is relatively high.

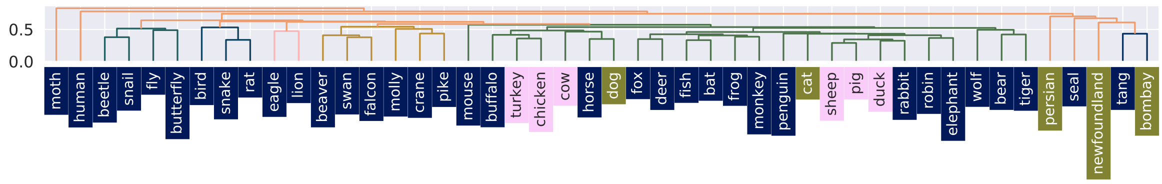
For the corpus-based bias evaluation, we replace relative pronouns referring to animal names with [MASK] tokens in extracted sentences. Then, we



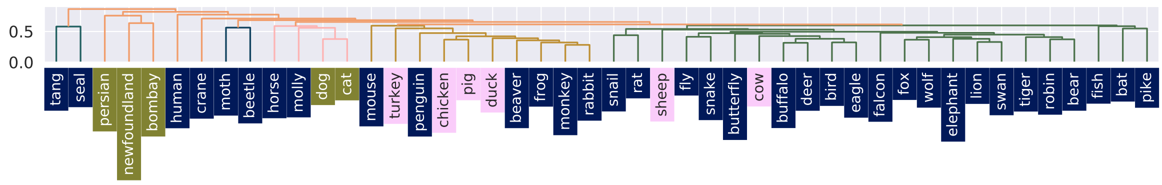
(a) BERT



(b) RoBERTa



(c) DistilBERT



(d) ALBERT

Figure 1: Results of hierarchical clustering based on the agreement rate of words whose predicted probability of filling the [MASK] token changed significantly between template sentences. Each leaf is colored using SciPy library (Virtanen et al., 2020), with the default color threshold.

use MLMs to calculate probabilities of relative pronouns at the [MASK] token. We compare the probabilities for both sets and evaluate the bias as follows:

$$\begin{aligned} bias = &\frac{1}{|H|}\sum_{s_i \in H}^{|H|} \mathbb{1}[p_{object|s_i} > p_{human|s_i}] \\ &- \frac{1}{|O|}\sum_{s_j \in O}^{|O|} \mathbb{1}[p_{human|s_j} > p_{object|s_j}] \end{aligned} \tag{3}$$

where H and O are the sets of human and object sentences extracted from Books3, and $s_{i,j}$ is a given sentence. $\mathbb{1}[\cdot]$ returns 1 if its condition is true and 0 otherwise. $p_{object|s_i}$ and $p_{human|s_i}$ are represented as follows:

$$\begin{aligned} p_{object|s_i} &= \max(p_{that|s_i}, p_{which|s_i}) \\ p_{human|s_i} &= \max(p_{who|s_i}, p_{whose|s_i}, p_{whom|s_i}) \end{aligned}$$

Variables $p_{that|s_i}$, $p_{which|s_i}$, $p_{who|s_i}$, $p_{whose|s_i}$, and $p_{whom|s_i}$ are the probabilities of each relative pronoun substituting [MASK] in a given sentence. If the value of the first term in the Equation 3 is closer to 1, MLMs incorrectly predict higher probability of "which" or "that", and if the second term approaches 1, MLMs incorrectly predict higher probability of "who", "whose" or "whom". In other words, when the bias is close to 1, models tend to regard animals as objects; and if it is close to -1, they tend to treat them as humans.

To investigate the relationship between the bias represented in Equation 3 and the frequency bias in the corpora, we also calculate the correlation between the bias and the frequency of object-related pronouns ("that" and "which") referring to each animal name in Wikipedia and BookCorpus.

5 Experimental Results

5.1 Template-based Evaluation

5.1.1 Probability Differences

The experimental results of probability differences between human and object sentences are presented

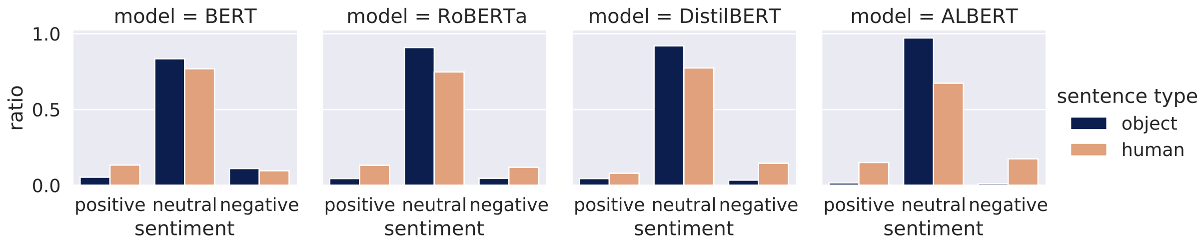


Figure 2: Results of sentiment analysis for each language model. Vertical axis shows the ratio of words assigned to a certain sentiment. For each sentiment, the darker bars indicate the percentage of words that have a higher mean probability in the object sentences, and the light-colored ones show the ratio of words that have a higher mean probability in the human sentences.

Table 2: Sets of five predicted words with the highest change rate in BERT. Possibly harmful biased words are shown in **bold** font.

Animal name	Words with high probability change in *object* sentences	Words with high probability change in *human* sentences
cat	**f**ked**, **f**king**, **reproduced**, violated, **ripe**	**sarcastic**, **mute**, Ninja, **clumsy**, unnamed
dog	**f**ked**, **f**king**, struck, violated, committed	**sarcastic**, Ninja, **mute**, bisexual, unnamed
chicken	**slaughtered**, **f**ked**, **stamped**, **reproduced**, **ripe**	**clumsy**, **mute**, **sarcastic**, psychic, superhero
pig	**f**ked**, **stamped**, **slaughtered**, **reproduced**, **sin**	**clumsy**, **sarcastic**, **mute**, cheerful, blonde
turkey	**stamped**, **slaughtered**, **beef**, **ripe**, viable	**mute**, **clumsy**, psychic, **sarcastic**, **deaf**
fish	endemic, predatory, widespread, perennial, barred	heroine, **sarcastic**, Cinderella, princess, cheerful
fox	**f**ked**, happening, waking, calling, ours	**mute**, **sarcastic**, blonde, bisexual, **clumsy**
horse	**f**ked**, **sin**, violated, **stamped**, **ripe**	unnamed, pink, **sarcastic**, blonde, Ariel
human	ourselves, worth, ours, yours, our	bisexual, Ninja, **sarcastic**, blonde, lesbian

Table 3: Sets of five predicted words with the highest change rate in RoBERTa. Possibly harmful biased words are shown in **bold** font.

Animal name	Words with high probability change in *object* sentences	Words with high probability change in *human* sentences
cat	terrestrial, armoured, netted, scaled, predatory	foster, **deaf**, Transgender, Blind, Polish
dog	terrestrial, itself, predatory, defined, armoured	**deaf**, transsexual, foster, Homeless, lesbian
chicken	**dried**, freshwater, semen, **polled**, **harvested**	optimistic, **sarcastic**, romantic, pessimistic, Psychic
pig	**polled**, **dried**, **harvested**, yielded, **peeled**	romantic, **selfish**, optimistic, jealous, **arrogant**
turkey	**dried**, **processed**, ground, **slaughtered**, cached,	**deaf**, listening, jealous, optimistic, psychic
fish	freshwater, reef, widespread, **polled**, aggregate	**swearing**, jealous, witty, superhuman, sixteen
fox	**polled**, invasive, Madagascar, pictured, extant	pessimistic, **sarcastic**, **mercenary**, romantic, compassionate
horse	clicking, enough, beat, it, right	Transgender, lesbian, **deaf**, transgender, transsexual
human	extant, extinct, ours, yours, **edible**	bartender, nineteen, seventeen, sixteen, eighteen

in Figure 1, and Figures 4, 5, 7, 6 in Appendix A.

From Figure 1 it can be observed that the names of animals colored with the same color belong to roughly the same clusters. Especially in the results of BERT and RoBERTa, the names of animals who are often kept at farms were clustered closely in most cases (see Figures 1a and 1b). In the results of DistilBERT and ALBERT, the animal names with the same color were not grouped together, but some belonged to the same cluster, indicating that they were not completely disjointed.

In Tables 2 and 3, we show sets of top five words with the largest probability change for each animal. For these tables we chose the five most frequent animal names in Wikipedia, and added the most popular animals living in farms and at homes, as they are one of the focal interest of our investigation: “cat”, “dog”, “chicken” and “pig”. In these tables, we show the results for BERT and RoBERTa, while the results of the remaining models are given in Appendix A.

For “chicken”, “pig” and “turkey”, words with high probability change in object sentences included “slaughtered”, “reproduced”, “ripe” (see Table 2), also “dried” and “harvested” (see Table

Table 4: Frequency of relative pronouns referring to animal names in each corpus (references determined by CoreNLP).

Corpus	that	which	who	whose	whom
Books3	104,244 (103,361)	28,552 (28,231)	44,607 (39,593)	4,115 (4,012)	2,006 (1,690)
Books-Corpus	5,111 (4,949)	1,470 (1,419)	3,925 (2,988)	183 (171)	66 (50)
Wikipedia (EN)	9,341 (9,265)	6,642 (6,586)	7,182 (6,648)	411 (396)	289 (274)

Table 5: Pearson correlation coefficient (r) between the bias represented in Equation 3 and frequency of object-related pronouns in Wikipedia and BookCorpus.

	BERT	RoBERTa	DistilBERT	ALBERT
r	0.77	0.55	0.81	0.74

3). Also, in BERT, "f**k"-rooted words were associated with many animals. On the other hand, in human sentences, associated words express personality and gender-related attributes, such as "clumsy" or "bisexual". There are also many words that represent personality traits that can be interpreted as negative, for example "sarcastic". However, "human" does not exhibit many such characteristics.

5.1.2 Sentiment Analysis

Next, we report the results of the sentiment analysis performed on each cluster obtained in the experiment described in 4.1.2 (see Figure 2). The vertical axis of the figure shows the percentage of the number of words assigned to each sentiment. The horizontal one shows the sentiment and the names of the models.

We found that VADER assigned 0 (i.e. neutral sentiment) to the majority of the words, and that object sentences contained more neutral words than human sentences in all models. Contrary to our hypothesis, the ratio of negative words was found to be larger in human sentences for all three models except BERT. Within each model, the distribution of assigned sentiment was generally the same.

5.2 Corpus-based Evaluation

Here, we present the results of the corpus-based experiment. First, we look at the sentences extracted from the corpora. In Table 4 we show the total number of relative pronouns referring to animal names in each corpus. The number in brackets indicates the total number minus the number of relative pronouns referring to "human". The total number for each animal is shown in Figures 8, 9 and 10. Comparing the total number of "that" and "which" with the total number of "who", "whose" and "whom", we found that the former is about twice more common. This indicates that the corpus as a whole tends to treat nonhuman animals as objects. In addition, contrary to our assumption, the number of relative pronouns such as "who" that refers to "dogs" and "cats" in all corpora is almost the same as the total number of "that" and "which" (see Figures 8 and 9).

Next, we examine the results of analyzing the bias of MLMs using sentences collected from the Books3 corpus (see Figure 3). The vertical axis of each graph represents the degree of bias, and the horizontal one represents the animal names. A positive bias indicates a high probability of incorrectly entering "that" or "which" (i.e., having a speciesist bias), while a negative bias indicates a high probability of incorrectly filling "who", "which", or "which" (i.e., having a non-speciesist bias).

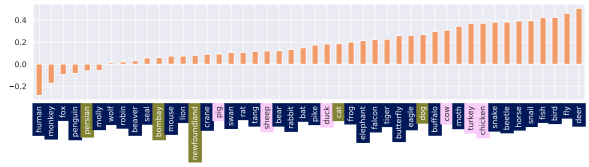
All of the models exhibited a negative bias against "human", and a positive bias against "chicken" and "turkey". These results are in line with our expectations. However, contrary to our predictions, the bias for "dog" and "cat" in BERT and RoBERTa is positive, indicating that they tend to be treated as objects. On the other hand, DistilBERT and ALBERT were found to include more negative bias, i.e. non-speciesist tendency, compared to BERT and RoBERTa. Table 5 shows the correlation between these biases and the ratio of the frequency of object-related pronouns in the corpora. The correlation was above 0.7 for MLMs other than RoBERTa, and above 0.5 for RoBERTa, which indicates that the ratio of relative pronouns in the corpus explains the bias of MLMs to some extent. We think that the low value for RoBERTa is due to the fact that RoBERTa has been pre-trained on other corpora.

6 Discussion

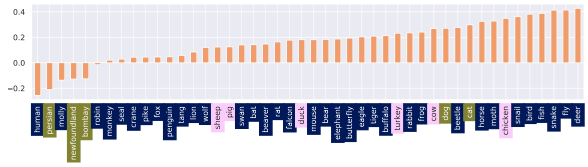
6.1 Template-based Approach

The results of the animal names clustering in BERT and RoBERTa partially support our hypothesis, which indicates that these models alter the words associated with animals between object and human sentences. On the other hand, DistilBERT and ALBERT performed clustering slightly different from our expectation, which may be due to the lower performance of mask predictions caused by the smaller model size.

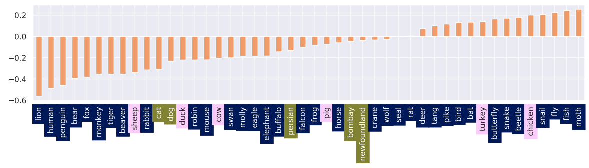
As shown in Tables 2 and 3, when nonhuman



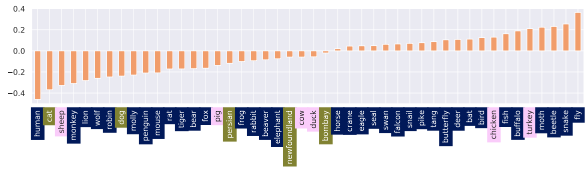
(a) BERT



(b) RoBERTa



(c) DistilBERT



(d) ALBERT

Figure 3: Results of the corpus-based bias analysis, sorted by the magnitude of the bias represented by Equation 3. Vertical axis shows the magnitude of the bias, where positive values indicate that MLMs incorrectly insert “that” or “which”, and negative values indicate that MLMs incorrectly insert “who”, “whose”, or “whom” with higher probability. The horizontal one shows the animal names. Bigger versions of the graphs are given in Appendix A.

animals are described by object sentences, they are linked with harmful words such as “f**ked”. Furthermore, in the case of animals who live in farms to be utilized as flesh, meat-related words have been confirmed, for example “slaughtered” and “harvested” described as problematic in previous studies (Dunayer, 2001, 2003, see also Section 2.2). These words are likely to be associated with speciesist language that objectifies animals.

In the experiments of sentiment analysis, it is important to note here that VADER itself may exhibit a speciesist bias. For example, VADER considers “killed” to be a negative word, but recognizes “slaughtered” as a neutral word. This problem should be investigated further.

6.2 Corpus-based Approach

Frequencies of human-related pronouns are lower than object-related pronouns in all corpora (see Table 4). There are at least two possible causes for this discrepancy: (1) there are fewer human-related relative pronouns that refer to nonhuman animals in the corpus than object-related ones, or (2) the recall of CoreNLP for human-related relative pronouns is low. If (1) is correct, it suggests that people tend to treat nonhuman animals as objects. If (2) is correct, it suggests that there is a bias in CoreNLP which makes the parser unable to sufficiently capture human-related relational references to nonhuman animals. Either result could be indirectly harmful to nonhuman animals.

In our corpus bias evaluation experiments, we found that, contrary to our hypothesis, the models had a speciesist bias against “dog” and “cat”. However, all models exhibited a non-speciesist bias for more specific kinds of dogs and cats such as “newhoundland” and “persian”. These results suggest that MLMs predicted “that” and “which” referring to “dog” and “cat” with high probability because they are commonly used as general names and therefore do not represent specific individuals. The bias between general names and more specific names will also be a subject of our future work.

7 Conclusion

In this paper, we analyze the speciesist bias against animals inherent in MLMs. Our experimental results show that such models strongly associate harmful words with many nonhuman animals. We also found that MLMs, especially BERT and RoBERTa, are biased to associate object-related pronouns (“that” and “which”) with various nonhuman animals, and demonstrate that this bias is correlated with the frequency of these relative pronouns referring to each animal in the corpora.

Since this research is restricted to English language, it cannot be generalized to other languages. Moreover, this paper does not address so-called *intersectional bias*. For example, “bitch” means a female dog, but it is also used as an insult toward women. In future, we plan to expand our research by utilizing findings in animal ethics regarding intersectional bias and discrimination between speciesist bias and other biases (Birke et al., 1995; Adams, 1990).

References

Carol J Adams. 1990. *The sexual politics of meat: A feminist-vegetarian critical theory*. Bloomsbury Publishing USA.

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.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Lynda Birke, Joan Dunayer, and Marti Kheel. 1995. *Animals and women: Feminist theoretical explorations*. Duke University Press.

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.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Fabio Crameri. 2021. Scientific colour maps.

Fabio Crameri, Grace E Shephard, and Philip J Heron. 2020. The misuse of colour in science communication. *Nature communications*, 11(1):1–10.

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.

Joan Dunayer. 1995. Sexist words, speciesist roots. In *Animals and women: Feminist theoretical explorations*, pages 11–31. Duke University Press Durham, NC.

Joan Dunayer. 2001. *Animal Equality: Language and Liberation*. Ryce Pub.

Joan Dunayer. 2003. English and speciesism. *English Today*, 19(1):61–62.

Emma Franklin. 2020. *Acts of killing, acts of meaning: an application of corpus pattern analysis to language of animal-killing*. Ph.D. thesis, Lancaster University.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *Computing Research Repository*, arXiv:2006.03955.

Deborah Hellman. 2008. *When is discrimination wrong?* Harvard University Press.

Oscar Horta and Frauke Albersmeier. 2020. Defining speciesism. *Philosophy Compass*, 15(11):e12708.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 5491–5501, Online. Association for Computational Linguistics.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Jill Jepson. 2008. A linguistic analysis of discourse on the killing of nonhuman animals. *Society & Animals*, 16(2):127–148.

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, 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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Anne Lauscher, Tobias Lüken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Charles D Michener and Robert R Sokal. 1957. A quantitative approach to a problem in classification. *Evolution*, 11(2):130–162.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, 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.

Andrea Owe and Seth D Baum. 2021. Moral consideration of nonhumans in the ethics of artificial intelligence. *AI and Ethics*, pages 1–12.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Shawn Presser. 2020. Books3 https://twitter.com/theshawwn/status/1320282149329784833.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. 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 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Alison Sealey and Chris Pak. 2018. First catch your corpus: methodological challenges in constructing a thematic corpus. *Corpora*, 13(2):229–254.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 4275–4293, Online. 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.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Peter Singer. 2015. *Animal Liberation*. Vintage Digital.

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.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Yi Chern Tan and L Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems*, volume 32, pages 13230–13241. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *Computing Research Repository*, arXiv:2010.06032.

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.

A Appendix

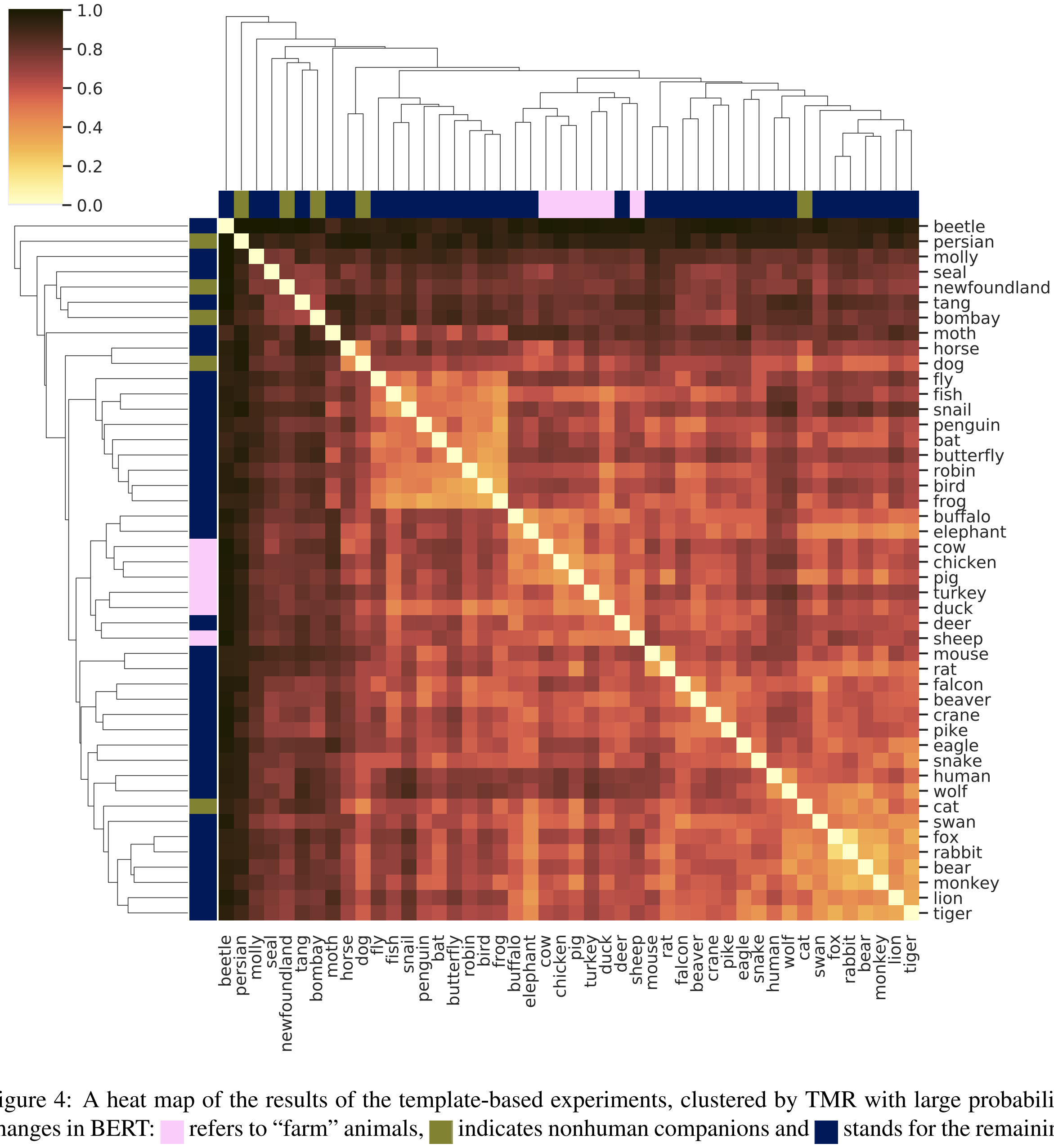


Figure 4: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in BERT: refers to "farm" animals, indicates nonhuman companions and stands for the remaining species.

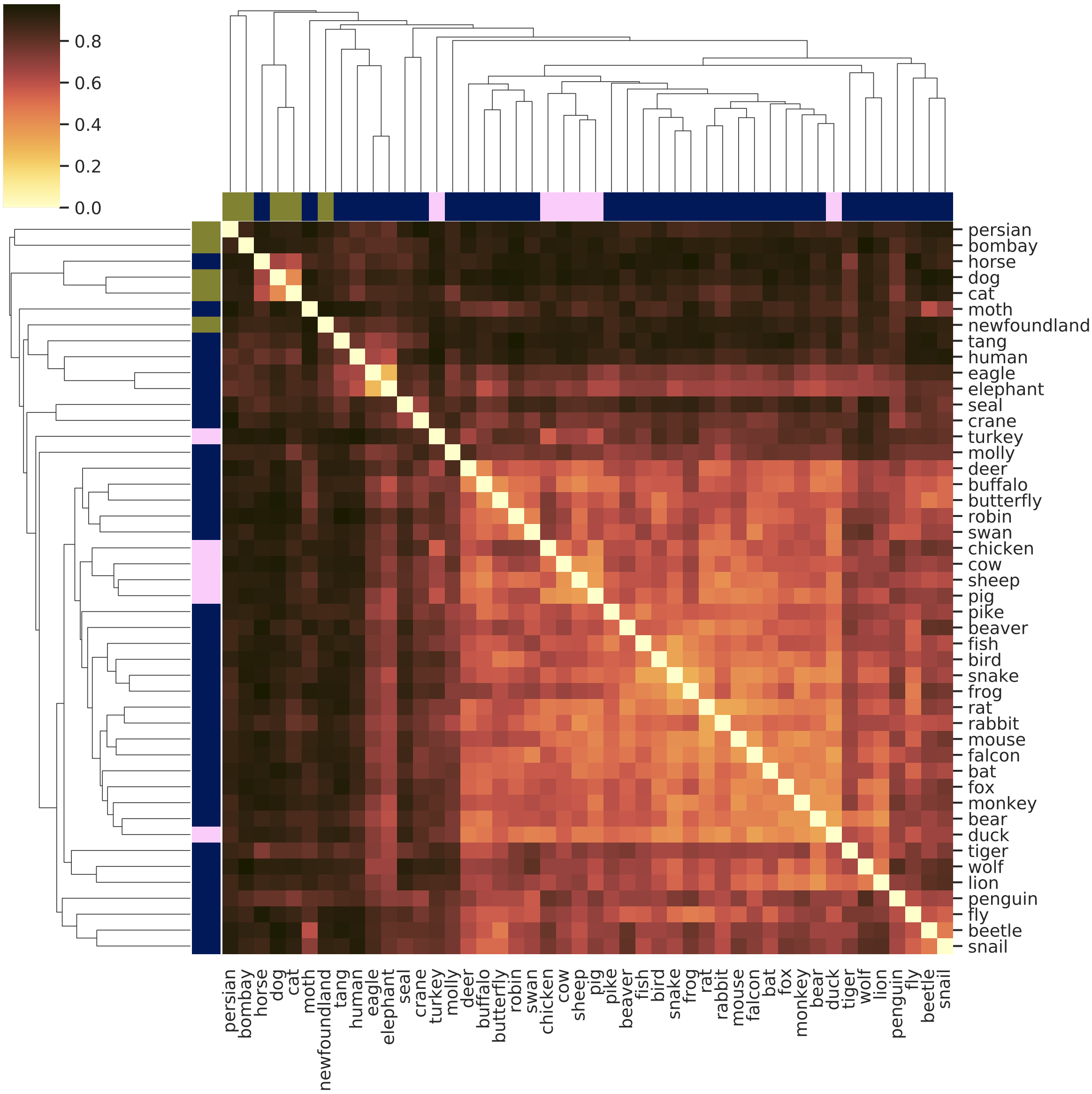


Figure 5: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in RoBERTa.

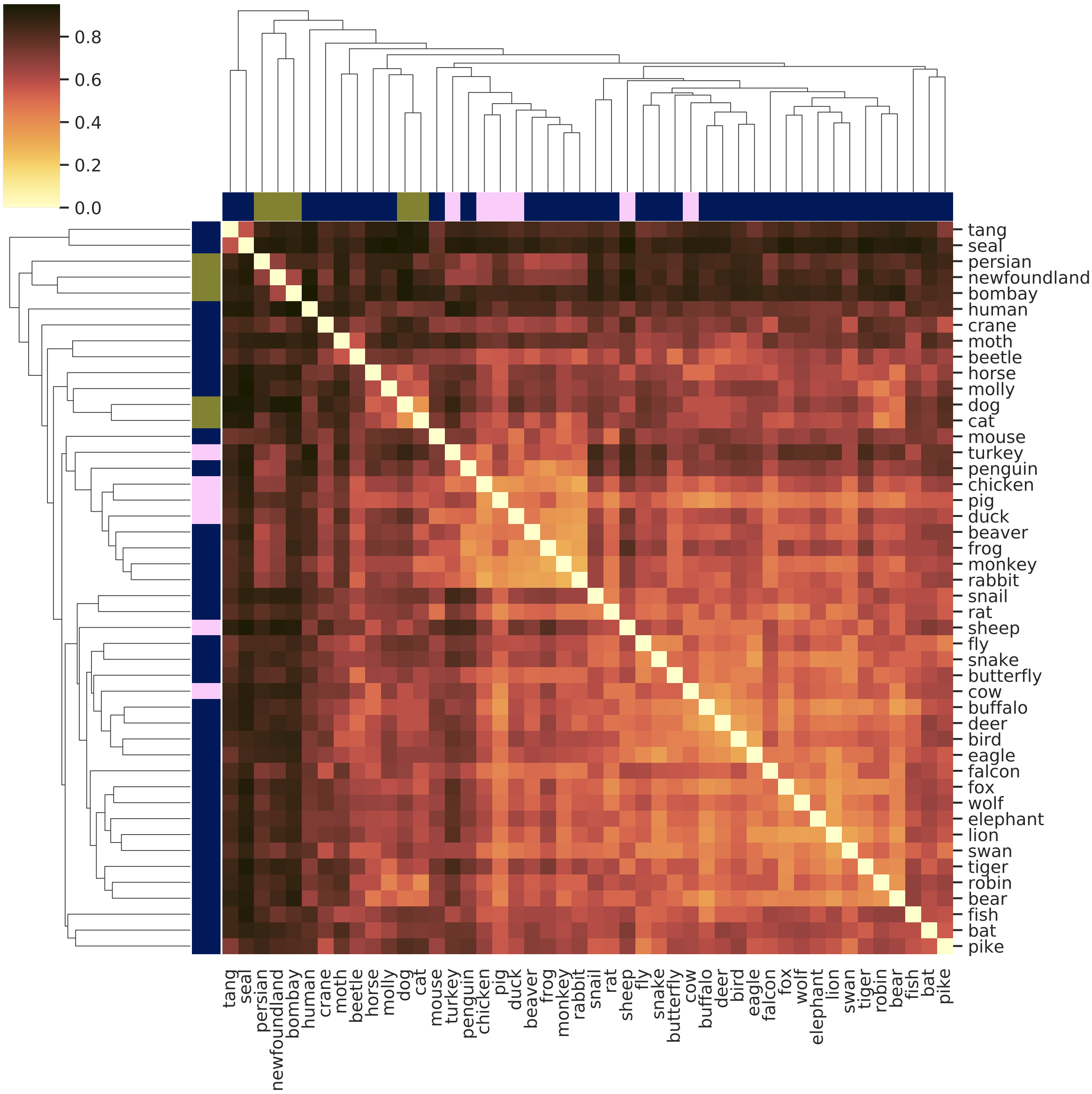


Figure 6: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in ALBERT

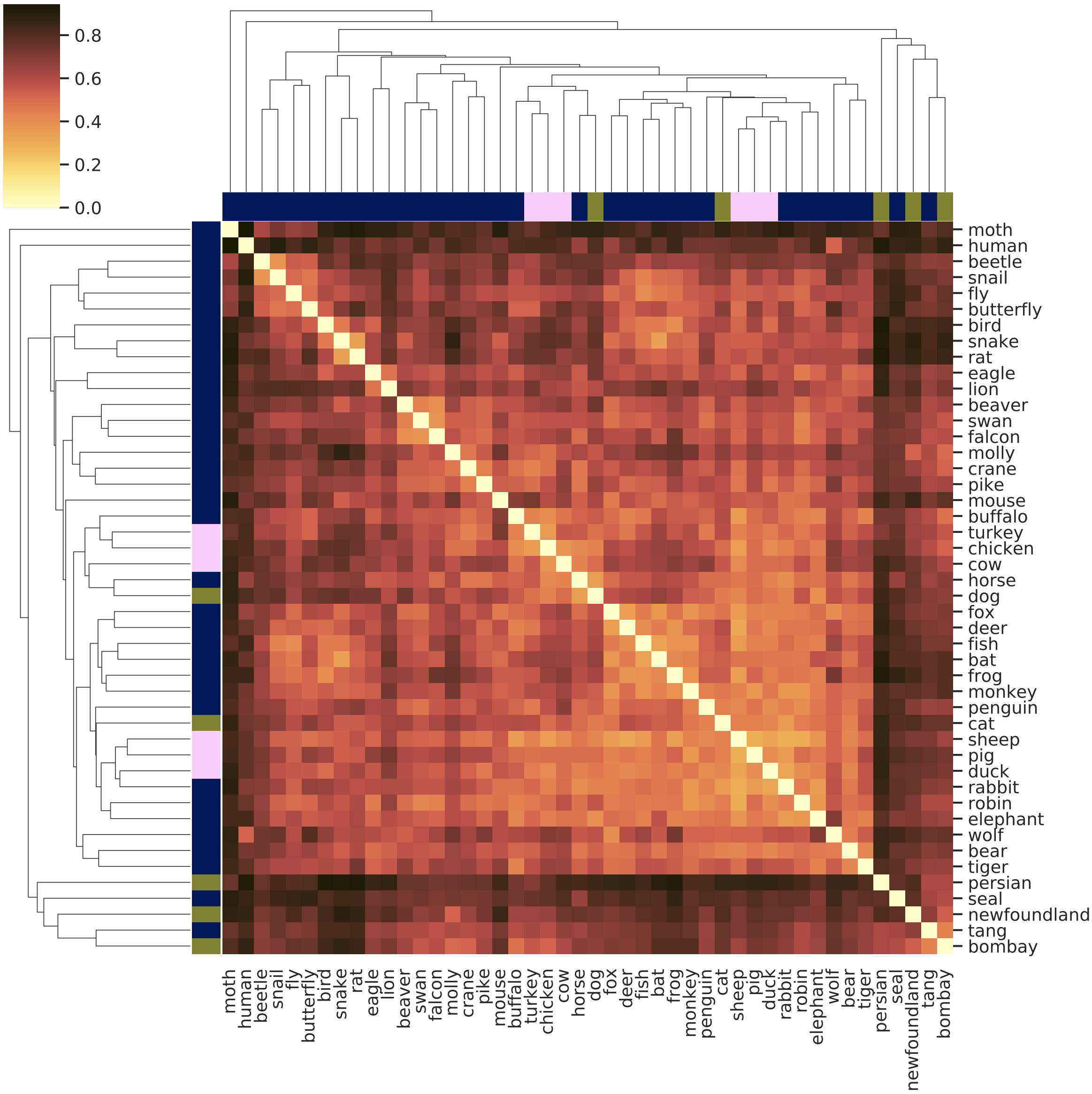


Figure 7: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in DistilBERT

Table 6: Sets of five words with the highest change rate in BERT

animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	endemic, threatened, predatory, barred, endangered	Ninja, sarcastic, blonde, Nordic, psychic
bear	f**ked, ours, waking, happening, stirring	sarcastic, bisexual, mute, psychic, blonde
beaver	endemic, reproduced, f**ked, extinct, viable	mute, psychic, Ninja, sarcastic, superhero
beetle	conspicuous, stinging, waking, ripe, variable	coach, coaching, coaches, Swiss, midfielder
bird	endemic, threatened, uncommon, endangered, widespread	sarcastic, blonde, psychic, superhero, heroine
bombay	standardized, portable, timed, ceremonial, audible	unemployed, widowed, homeless, heroine, psychologist
buffalo	stamped, f**ked, beef, slaughtered, reproduced	psychic, mute, sarcastic, clumsy, blind
butterfly	endemic, widespread, uncommon, threatened, disputed	sarcastic, superhero, blonde, cheerful, mute
cat	f**ked, f**king, reproduced, violated, ripe	sarcastic, mute, Ninja, clumsy, unnamed
chicken	slaughtered, f**ked, stamped, reproduced, ripe	clumsy, mute, sarcastic, psychic, superhero
cow	f**ked, stamped, slaughter, slaughtered, ripe	sarcastic, mute, clumsy, psychic, cheerful
crane	loading, rotating, operating, overhead, tuned	psychic, blonde, sarcastic, mute, heroine
deer	beef, f**ked, f**king, viable, barred	mute, fairy, sarcastic, Ariel, psychic
dog	f**ked, f**king, struck, violated, committed	sarcastic, Ninja, mute, bisexual, unnamed
duck	endemic, f**ked, reproduced, endangered, viable	sarcastic, psychic, clumsy, mute, cheerful
eagle	barred, circling, happening, endemic, yours	mute, psychic, sarcastic, Amazon, blonde
elephant	f**ked, stamped, reproduced, f**king, happening	sarcastic, mute, psychic, clumsy, cheerful
falcon	barred, endemic, f**ked, reproduced, extinct	blonde, Ninja, psychic, sarcastic, Amazon
fish	endemic, predatory, widespread, perennial, barred	heroine, sarcastic, Cinderella, princess, cheerful
fly	predatory, toxic, stinging, endemic, colonial	sarcastic, cheerful, superhero, blonde, genius
fox	f**ked, happening, waking, calling, ours	mute, sarcastic, blonde, bisexual, clumsy
frog	endemic, threatened, endangered, ##olate, widespread	sarcastic, Ninja, cheerful, clumsy, blonde
horse	f**ked, sin, violated, stamped, ripe	unnamed, pink, sarcastic, blonde, Ariel
human	ourselves, worth, ours, yours, our	bisexual, Ninja, sarcastic, blonde, lesbian
lion	ours, happening, waking, arising, pictured	psychic, sarcastic, mute, heroine, bisexual
molly	unacceptable, theirs, treason, happening, occurring	blonde, mute, deaf, cheerful, widowed
monkey	f**ked, f**king, waking, happening, ours	sarcastic, mute, clumsy, lesbian, Ninja
moth	endemic, Crambidae, ##tropical, variable, Geometridae	unemployed, Ninja, DJ, psychic, undefeated
mouse	reproduced, viable, mating, f**ked, endemic	sarcastic, cheerful, clumsy, Dorothy, superhero
newfoundland	ours, theirs, happening, paradise, nearer	bisexual, protagonist, narrator, heroine, blonde
penguin	endemic, extinct, endangered, barred, reproduced	sarcastic, psychic, Ninja, clumsy, mute
persian	periodic, convex, contraction, symmetric, bounded	deaf, genius, widowed, ##headed, intelligent
pig	f**ked, stamped, slaughtered, reproduced, sin	clumsy, sarcastic, mute, cheerful, blonde
pike	endemic, barred, preferred, edged, subspecies	blonde, mute, widowed, cheerful, homeless
rabbit	f**ked, waking, happening, slaughtered, arriving	mute, sarcastic, bisexual, psychic, clumsy
rat	reproduced, f**ked, viable, reared, waking	sarcastic, mute, clumsy, Gothic, cheerful
robin	endemic, subspecies, threatened, barred, unmistakable	mute, sarcastic, psychic, cheerful, mechanic
seal	stamped, forged, valid, void, binding	Brave, blonde, Ninja, psychic, mute
sheep	endemic, f**ked, sustainable, perennial, viable	mute, sarcastic, psychic, princess, narrator
snail	predatory, endemic, widespread, fossil, marine	sarcastic, cheerful, mute, optimistic, psychic
snake	endemic, yours, barred, ours, venom	sarcastic, cheerful, blonde, mute, optimistic
swan	yours, ours, f**ked, reproduced, endemic	psychic, sarcastic, mute, mechanic, clumsy
tang	audible, repeated, nasal, consonant, pronounced	Smart, unemployed, smart, homeless, brave
tiger	happening, ours, f**king, waking, f**ked	mute, sarcastic, psychic, bisexual, blonde
turkey	stamped, slaughtered, beef, ripe, viable	mute, clumsy, psychic, sarcastic, deaf
wolf	ours, yours, happening, waking, you	bisexual, mute, sarcastic, psychic, lesbian

Table 7: Sets of five words with the highest change rate in RoBERTa

animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	intact, handled, dried, unloaded, batted	virtuous, heroic, witty, superhuman, princess
bear	polled, extant, freshwater, handled, endemic	superhuman, mercenary, romantic, sarcastic, prince
beaver	invasive, freshwater, widespread, dried, common	atheist, lonely, swearing, nineteen, lesbian
beetle	deposited, feeding, circulating, hardest, clustered	virtuous, heroic, fictional, philosophical, courageous
bird	freshwater, offshore, migr, endemic, extant	Human, philosophical, jealous, sarcastic, witty
bombay	fallacy, phosphorus, absurdity, gelatin, FALSE	Shy, loyal, married, shy, wealthy
buffalo	dried, freshwater, listed, polled, stamped	cowardly, arrogant, selfish, cunning, rebellious
butterfly	variable, common, offshore, widespread, clustered	virtuous, superhuman, philosophical, rebellious, heroic
cat	terrestrial, armoured, netted, scaled, predatory	foster, deaf, Transgender, Blind, Polish
chicken	dried, freshwater, semen, polled, harvested	optimistic, sarcastic, romantic, pessimistic, Psychic
cow	polled, dried, semen, domestically, processed	romantic, optimistic, witty, poetic, mysterious
crane	erected, automated, propelled, loader, towed	jealous, psychic, horny, deaf, conflicted
deer	bucks, dried, harvested, roadside, buck	swearing, witty, romantic, philosophical, jealous
dog	terrestrial, itself, predatory, defined, armoured	deaf, transsexual, foster, Homeless, lesbian
duck	freshwater, polled, dried, offshore, netted	superhuman, heroic, superhero, protagonist, Human
eagle	correlated, achievable, warranted, measurable, irreversible	adventurer, hacker, Paladin, Sailor, trainer
elephant	achievable, warranted, happening, extinct, irreversible	adventurer, detective, Lesbian, thief, vigilante
falcon	freshwater, netted, largest, aerial, perched	Human, optimistic, superhuman, rebellious, lesbian
fish	freshwater, reef, widespread, polled, aggregate	swearing, jealous, witty, superhuman, sixteen
fly	respiratory, common, genital, dried, larvae	heroic, lonely, witty, Talking, intuitive
fox	polled, invasive, Madagascar, pictured, extant	pessimistic, sarcastic, mercenary, romantic, compassionate
frog	freshwater, larvae, widespread, invasive, dart	superhuman, seventeen, nineteen, swearing, heroic
horse	clicking, enough, beat, it, right	Transgender, lesbian, deaf, transgender, transsexual
human	extant, extinct, ours, yours, edible	bartender, nineteen, seventeen, sixteen, eighteen
lion	pictured, Madagascar, Guinea, polled, Bengal	Human, prince, princess, mercenary, Princess
molly	edible, larvae, harvested, dried, invasive	pessimistic, Persian, nineteen, deaf, lazy
monkey	polled, palm, Madagascar, extant, Guinea	virtuous, mercenary, superhuman, Alone, romantic
moth	happening, circulating, newer, collapsing, getting	prophetic, divine, :, Blind, feminist
mouse	polled, larvae, extant, freshwater, edible	swearing, romantic, Alone, heroic, rich
newfoundland	Antarctica, unfolding, contiguous, ours, wetlands	deaf, transsexual, bisexual, runner, addicted
penguin	lower, offshore, flattened, freshwater, oval	lesbian, unmarried, married, rebellious, feminist
persian	larvae, edible, peeled, citrus, vegetation	atheist, writer, novelist, journalist, physicist
pig	polled, dried, harvested, yielded, peeled	romantic, selfish, optimistic, jealous, arrogant
pike	freshwater, offshore, invasive, Atlantic, harvested	Human, protector, nineteen, optimistic, swearing
rabbit	dried, widespread, netted, terrestrial, harvested	sarcastic, Psychic, optimistic, pessimistic, heroic
rat	dried, freshwater, widespread, polled, extant	heroic, swearing, superhuman, romantic, protector
robin	common, variable, migrating, widespread, larvae	superhuman, virtuous, philosophical, trustworthy, irresponsible
seal	tightening, tightened, tighter, stamped, dried	autistic, Hungry, dreaming, deaf, transsexual
sheep	polled, dried, harvested, yielded, processed	jealous, witty, arrogant, heroic, optimistic
snail	minute, deposited, dried, flattened, occurring	clueless, Psychic, jealous, cowardly, loyal
snake	freshwater, netted, dried, invasive, widespread	superhuman, swearing, cursed, immortal, protagonist
swan	freshwater, aerial, lower, largest, netted	protector, trustworthy, forgiving, pessimistic, loyal
tang	contraction, residue, correlation, causation, correlated	deaf, homeless, transsexual, Homeless, veterinarian
tiger	manageable, corrected, viable, right, largest	lesbian, princess, transsexual, vegan, Human
turkey	dried, processed, ground, slaughtered, cached	deaf, listening, jealous, optimistic, psychic
wolf	polled, extant, heaviest, widespread, invasive	Psychic, wizard, Human, Loki, prince

Table 8: Sets of five words with the highest change rate in DistilBERT

animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	endemic, distributed, widespread, ##olate, ##gratory	magician, psychic, witch, villains, wizard
bear	endemic, distributed, valid, edible, convex	psychic, witches, witch, herself, grandmother
beaver	distributed, endemic, lateral, ##gratory, inactivated	psychic, heroine, archaeologist, magician, narrator
beetle	endemic, widespread, distributed, subsp, valid	magician, transgender, psychic, widowed, deaf
bird	endemic, distributed, widespread, variable, declining	robot, psychic, princess, witches, angel
bombay	quarterly, annual, administered, recited, yearly	widowed, transgender, deaf, bisexual, blind
buffalo	endemic, abolished, extinct, inactivated, edible	heroine, actress, girlfriend, psychic, narrator
butterfly	endemic, widespread, distributed, valid, decreasing	lion, psychic, controlling, gifted, vain
cat	endemic, valid, convex, inactivated, viable	narrator, thirteen, psychic, fourteen, seventeen
chicken	endemic, edible, pounded, differentiated, clarified	deaf, blind, psychic, narrator, bullying
cow	endemic, edible, differentiated, sacred, branched	homeless, deaf, bullying, blind, paranoid
crane	distributed, towed, valid, endemic, unfolded	psychic, heroine, deaf, magician, actress
deer	endemic, ##gratory, distributed, extinct, subspecies	psychic, sailor, narrator, witch, grandmother
dog	endemic, subspecies, branched, differentiated, valid	herself, teenage, widowed, thirteen, grandmother
duck	endemic, valid, distributed, subspecies, edible	psychic, narrator, clumsy, deaf, thirteen
eagle	endemic, distributed, valid, lateral, decreasing	psychic, princess, witches, herself, fairies
elephant	endemic, distributed, convex, valid, inhabited	heroine, magician, nurse, psychic, princess
falcon	convex, scaled, lateral, distributed, endemic	psychic, transgender, magician, controlling, kidnapped
fish	endemic, distributed, widespread, variable, diagnostic	widowed, narrator, sailor, genius, girlfriend
fly	distributed, valid, extant, endemic, occurring	deaf, motorcycle, sailor, narrator, thirteen
fox	endemic, distributed, ##gratory, extinct, extant	heroine, magician, psychic, sailor, narrator
frog	endemic, distributed, widespread, variable, valid	vain, princess, fairies, psychic, narrator
horse	valid, equivalent, propelled, endemic, assessed	heroine, grandmother, witches, fairies, princess
human	worth, acceptable, our, reproduced, valid	princess, witch, emerald, angel, witches
lion	endemic, engraved, displayed, seated, valid	psychic, heroine, princess, witches, witch
molly	frequented, underway, inhabited, unfinished, excavated	bisexual, deaf, transgender, elderly, widowed
monkey	endemic, distributed, valid, convex, differentiated	princess, witch, magician, herself, witches
moth	widespread, occurring, varies, irregular, subsp	blind, sighted, deaf, blinded, astronomer
mouse	distributed, endemic, inactivated, valid, bilateral	witches, fairies, thirteen, prostitutes, witch
newfoundland	endemic, populated, inhabited, frequented, dotted	widowed, secretary, bisexual, transgender, pregnant
penguin	endemic, valid, distributed, extinct, extant	psychic, widowed, narrator, magician, actress
persian	convex, bounded, periodic, continuous, compact	transgender, actress, widowed, nurse, wrestler
pig	endemic, edible, viable, differentiated, inactivated	thirteen, dolls, narrator, girlfriend, seventeen
pike	valid, longitudinal, endemic, distributed, convex	psychic, heroine, fairies, caring, narrator
rabbit	endemic, distributed, viable, differentiated, inactivated	magician, psychic, witch, witches, narrator
rat	endemic, distributed, oral, lateral, bilateral	witches, witch, fairies, wizard, princess
robin	endemic, distributed, valid, branched, widespread	psychic, heroine, autism, deaf, narrator
seal	stamped, filed, valid, worn, engraved	heroine, psychic, kidnapped, protagonist, drowning
sheep	endemic, distributed, inactivated, viable, extinct	psychic, housekeeper, narrator, thirteen, witch
snail	widespread, distributed, endemic, variable, minute	psychic, villain, widowed, protagonist, lion
snake	distributed, endemic, ##olate, variable, diagnostic	witch, princess, fairies, wizard, goddess
swan	endemic, distributed, ##tail, lateral, ##gratory	psychic, narrator, magician, transgender, autism
tang	recited, oral, cumulative, meaningful, elastic	blind, widowed, heroine, deaf, scientist
tiger	endemic, distributed, valid, extant, inhabited	heroine, widowed, princess, lovers, witch
turkey	endemic, extant, edible, widespread, valid	deaf, actress, psychic, transgender, narrator
wolf	endemic, valid, conspicuous, edible, variable	witches, witch, princess, fairies, grandmother

Table 9: Sets of five words with the highest change rate in ALBERT

animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	printed, basalt, lodged, cylindrical, mandible	confident, gambler, dreamer, fearless, grieving
bear	lodged, reported, indicated, suggested, excavated	trusting, helpless, fearless, trusted, obedient
beaver	noticeable, lodged, brownish, coughed, yellowish	heiress, dreamer, princess, bachelor, addict
beetle	leaked, brownish, lodged, occurring, yellowish	princess, adventurer, dreamer, valkyrie, knighted
bird	printed, brownish, yellowish, lodged, localized	dreamer, conqueror, princess, angels, slaves
bombay	reopened, commenced, redeveloped, expanded, skyline	widow, soprano, eunuch, knighted, pregnant
buffalo	brownish, lodged, yellowish, reported, basalt	dreamer, hero, helpless, obedient, widow
butterfly	printed, yellowish, brownish, highlighted, forewing	dreamer, conqueror, adventurer, himself, superhuman
cat	lodged, boar, appeared, urine, yellowish	dreamer, fearless, bachelor, confident, jed
chicken	spelt, brownish, lodged, compressed, stemmed	dreamer, atheist, jealous, telepathic, princess
cow	weigh, spelt, raked, brownish, lodged	destiny, conqueror, happiness, trusting, fearless
crane	corrugated, aluminium, hangar, diameter, turbine	jealous, dreamer, eunuch, homosexual, tigre
deer	brownish, lodged, reported, yellowish, surfaced	dreamer, slaves, conqueror, trusting, angels
dog	lodged, boar, suggested, reported, spelt	perfection, caring, fearless, loving, faithful
duck	spelt, contains, termed, spelled, containing	atheist, adventurer, dreamer, addict, estranged
eagle	resembled, printed, brachy, tapered, holotype	conqueror, helpless, widow, steward, dreamer
elephant	reported, brownish, yellowish, lodged, surfaced	helpless, conqueror, obedient, slaves, estranged
falcon	resembled, compressed, mandible, resembles, rectangular	dreamer, fearless, trusting, addict, obedient
fish	tapered, formulated, brownish, stemmed, tasted	dreamer, jealous, himself, atheist, conqueror
fly	nitrogen, printed, tapered, compressed, brownish	conqueror, dreamer, hostage, slaves, murderer
fox	brownish, yellowish, dorsal, bluish, puma	dreamer, trusted, slaves, trusting, selfish
frog	resembled, contains, spelt, termed, compressed	atheist, princess, bachelor, transgender, dreamer
horse	hoof, suggested, raked, overturned, hydraulic	caring, fearless, helpless, trusting, perfection
human	http, suggested, computed, spelt, stated	savior, loves, protector, loving, beloved
lion	noticeable, indicated, reported, yellowish, conical	trusting, estranged, helpless, dreamer, selfish
molly	spelled, suggested, advertised, yellowish, bacterio	confidant, dreamer, confident, obedient, fearless
monkey	resembled, spelt, xylo, termed, suggested	estranged, princess, dreamer, atheist, bachelor
moth	widespread, annual, biennial, localized, basal	dreamer, jed, magician, sorcerer, himself
mouse	generate, termed, contains, kernel, xml	wealthy, fearless, princess, billionaire, dreamer
newfoundland	happen, happened, place, reopened, resumed	widow, knighted, pregnant, transgender, addict
penguin	contained, brownish, noticeable, smelled, yellowish	widow, atheist, addict, billionaire, heiress
persian	quartz, sodium, clarified, indicated, contrary	heiress, wealthy, married, widow, unmarried
pig	brownish, termed, dorsal, yellowish, spelt	trusting, jealous, princess, selfish, estranged
pike	corrugated, diameter, tapered, aluminium, compressed	jealous, gambler, grieving, helpless, dreamer
rabbit	snout, resembled, contains, termed, spelt	dreamer, atheist, princess, transgender, estranged
rat	nitrogen, termed, contains, 1:, containing	dreamer, selfish, conqueror, estranged, atheist
robin	plumage, brownish, yellowish, printed, spelt	dreamer, confident, heroine, psychopath, selfish
seal	minimize, compress, tissue, membrane, corrugated	temeraire, racehorse, knighted, valkyrie, shepherd
sheep	aerobic, discontinued, uploaded, dorsal, reported	traitor, helpless, trusting, conqueror, slaves
snail	nitrogen, termed, corrugated, sodium, containing	selfish, dreamer, strangers, helpless, obedient
snake	localized, bluish, yellowish, pointed, brownish	messiah, conqueror, dreamer, helpless, obedient
swan	printed, tapered, erupted, conical, plumage	helpless, obedient, trusting, conqueror, estranged
tang	nitrogen, minimize, termed, compressed, compress	adventurer, conqueror, abbess, empress, barbarian
tiger	yellowish, brownish, bluish, excavated, reported	helpless, trusting, selfish, caretaker, incapable
turkey	tasted, dried, sliced, crisp, highlighted	adventurer, atheist, transgender, telepathic, knighted
wolf	mandible, dorsal, conical, termed, brownish	dreamer, helpless, princess, orphan, traitor

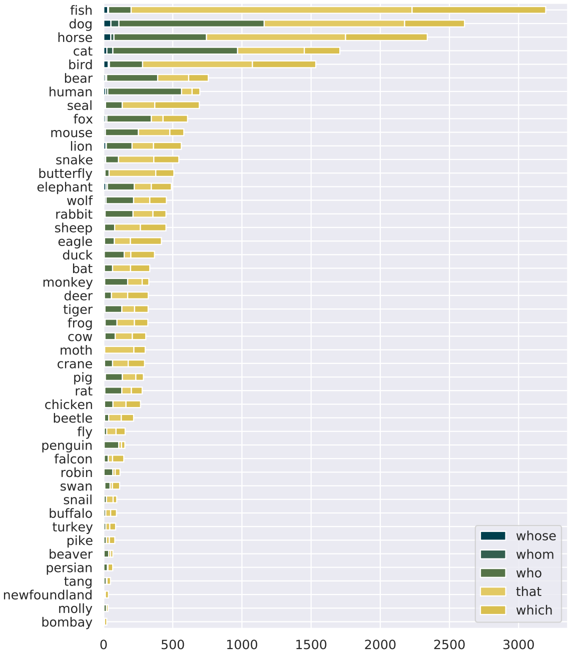


Figure 8: Number of relative pronouns referring to each animal in English Wikipedia.

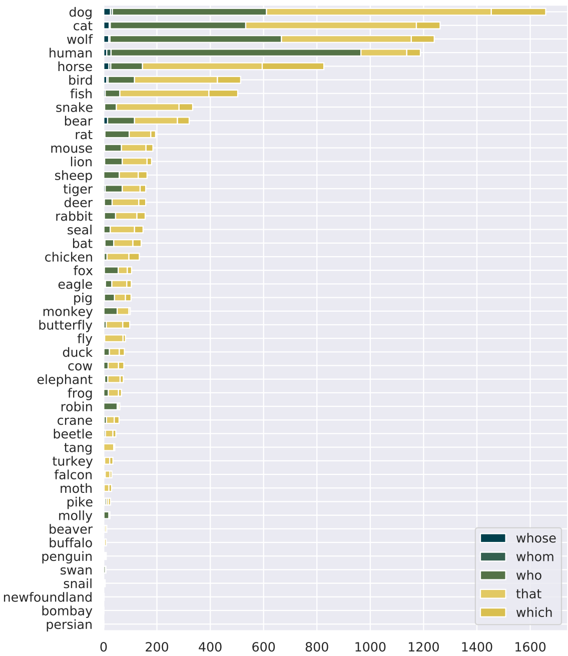


Figure 9: Number of relative pronouns referring to each animal in BookCorpus.

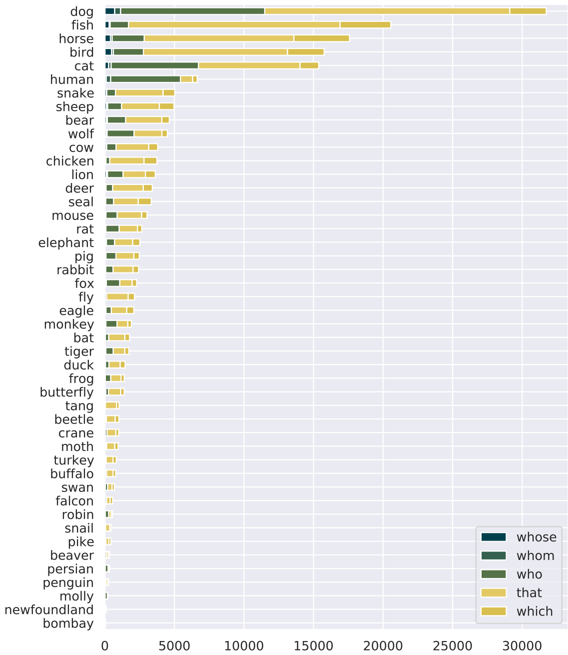
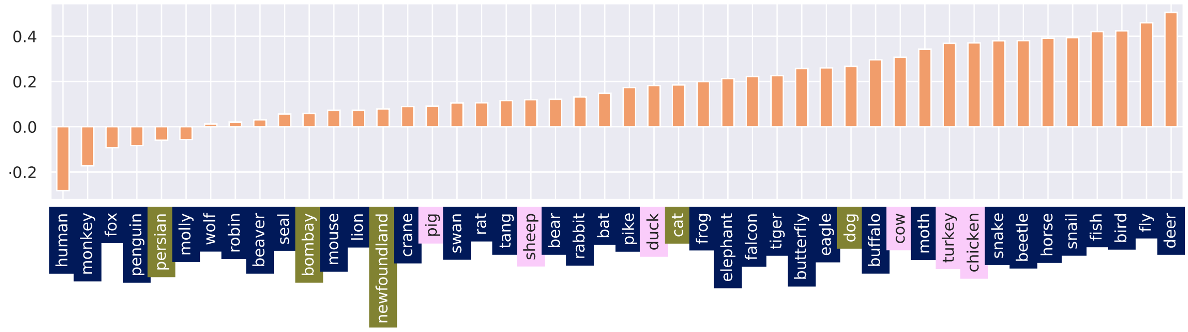
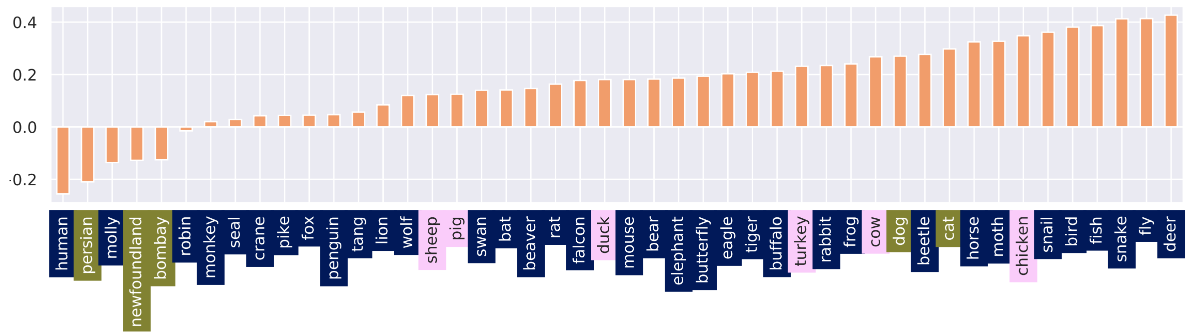


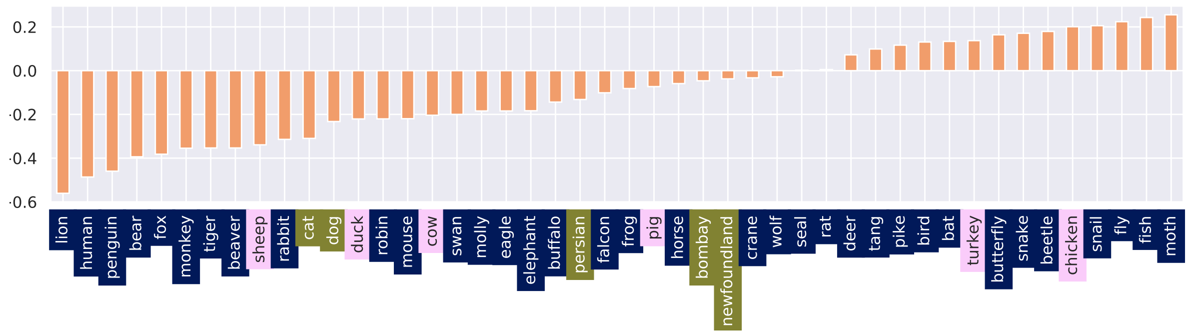
Figure 10: Number of relative pronouns referring to each animal in Books3.



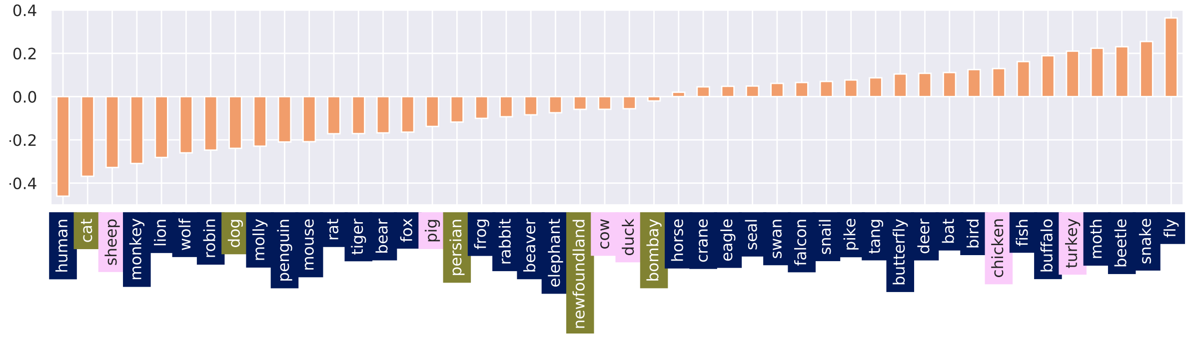
(a) BERT



(b) RoBERTa



(c) DistilBERT



(d) ALBERT

Figure 11: Results of the corpus-based bias analysis, sorted by the magnitude of the bias represented by Equation 3. Vertical axis shows the magnitude of the bias, where positive values indicate that MLMs incorrectly insert “that” or “which”, and negative values indicate that MLMs incorrectly insert “who”, “whose”, or “whom” with higher probability. The horizontal one shows the animal names.