Analyzing Occupational Distribution Representation in Japanese Language Models

Katsumi Ibaraki[1], **Winston Wu**[2], **Lu Wang**[1], **Rada Mihalcea**[1]
[1]University of Michigan
Ann Arbor, Michigan, USA
[2]University of Hawai'i at Hilo
Hilo, Hawai'i, USA
{ibaraki, wangluxy, mihalcea}@umich.edu
wswu@hawaii.edu

Abstract

Recent advances in large language models (LLMs) have enabled users to generate fluent and seemingly convincing text. However, these models have uneven performance in different languages, which is also associated with undesirable societal biases toward marginalized populations. Specifically, there is relatively little work on Japanese models, despite it being the thirteenth most widely spoken language. In this work, we first develop three Japanese language prompts to probe LLMs' understanding of Japanese names and their association between gender and occupations. We then evaluate a variety of English, multilingual, and Japanese models, correlating the models' outputs with occupation statistics from the Japanese Census Bureau from the last 100 years. Our findings indicate that models can associate Japanese names with the correct gendered occupations when using constrained decoding. However, with sampling or greedy decoding, Japanese language models have a preference for a small set of stereotypically gendered occupations, and multilingual models, though trained on Japanese, are not always able to understand Japanese prompts.

Keywords: Ethics and Legal Issues, Language Modeling, Multilinguality, Natural Language Generation

1. Introduction

With the recent advances in language generation, large language models (LLMs) are able to generate fluent and seemingly convincing text. However, these models can also produce text containing undesirable societal biases toward marginalized populations (Sheng et al., 2019; Wallace et al., 2019; Sheng et al., 2021). These biases are in part a result of the data that these models are trained on (Hovy and Prabhumoye, 2021; Gururangan et al., 2022). While much of the NLP community has focused on removing the association between occupation and gender (Blodgett et al., 2020), this idealized approach may not reflect real-world gender disparities (Touileb et al., 2022).

In this paper, we highlight the need for researchers to take these biases into account when working with Japanese language models, or more generally, LLMs. Additionally, we hope to assess how language-specific models reflect real-world distributions such as occupation. Though it is the 13th most widely-spoken language in the world, Japanese is relatively understudied in the NLP community. Only very recently have there been efforts to train and fine-tune LLMs for Japanese (Itoh and Shinnou, 2021; Yamauchi et al., 2022; Miyazaki et al., 2022; Ri et al., 2022).

We investigate the following research questions, focusing specifically on Japanese: How are gender and occupation represented in pre-trained language models? How are these distributions correlated with real-world statistics? How do models associate Japanese names with gender and occupation? To answer these questions, we probe language models' biases with respect to gender and occupations by developing natural language prompts in Japanese and measuring the differences in the LLMs' generated outputs with existing occupation statistics from the Japanese Census Bureau.

Japanese is an interesting language for investigation for several reasons. Models can be sensitive to morphosyntactic alignment, where word order and other grammatical features can affect the representations (Papadimitriou et al., 2021). As Japanese is a language with a Subject-Object-Verb (SOV) word order, models cannot be prompted the same way as more commonly studied languages with a Subject-Verb-Object (SVO) word order, such as English. Another challenge is that Japanese is a topic-prominent language, which emphasizes the topic-comment structure of a sentence, unlike English, which is a subject-prominent language that emphasizes the subject-predicate structure of a sentence (Li and Thompson, 1976). In subject-prominent languages, the subject (doer of the action) is placed first in a sentence, but with topic-prominent languages, the topic, which comes first in the sentence, may

LREC-COLING 2024, pages 959–973
20-25 May, 2024.

not necessarily be the sentence's subject. Finally, considering the gender differences, Japan is ranked 125th out of 146 nations by the World Economic Forum's global gender gap report.[1] If the Japanese models accurately represent Japan's cultural and societal aspects, we can expect substantial differences compared with multilingual and English models.

We evaluate models spanning a range of architectures and trained on various English, Japanese, or multilingual data. On name gender classification experiments, we find that Japanese-specific models trained with a masked language modeling objective performed most accurately. However, varying the prompts can cause other multilingual models to perform comparably to Japanese models. We additionally compare distributions of occupations from models and census statistics, and conduct temporal analyses of popular names and occupation statistics from the last 100 years, finding that models trained on Japanese Wikipedia tend to more closely match real-world occupation distributions.

2. Related Work

There has been much recent work that investigates bias in NLP models. For example, StereoSet (Nadeem et al., 2021) is a dataset designed to evaluate stereotypes in language models with regard to gender, profession, race, and religion. Dixon et al. (2018) measure and mitigate biases in toxicity classification models. There have been numerous metrics developed for measuring bias and fairness in NLP. Czarnowska et al. (2021) surveys these metrics. Sheng et al. (2021) is a survey of bias in language generation tasks. However, existing work has not investigated bias differences in models trained on Japanese.

Associations between occupations and gender have been investigated in coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), language modeling (Qian et al., 2019; Alnegheimish et al., 2022), sentiment analysis (Bhaskaran and Bhallamudi, 2019), and word embeddings (Caliskan et al., 2017). However, prior work focuses on English text and models. Our work is most similar to Touileb et al. (2022), who analyze occupations and gender in Norwegian language models. However, their work only investigates BERT-style masked language models. We also experiment with modern autoregressive models and other commercial LMs, and additionally conduct temporal analyses of names and occupations representative of different time periods.

[1] https://www.weforum.org/reports/global-gender-gap-report-2023/

3. Methods

Bias. We adopt an established definition of bias by Friedman and Nissenbaum (1996), who define bias as "systematically and unfairly discriminat[ion] against certain individuals or groups of individuals in favor of others." In language models, this bias may occur if skewed gender representations are not taken into account in downstream applications. In our work, we focus on the systematic association between occupations and gender (male/female) in language models. Rather than assuming that models should treat genders equally, we investigate how these models reflect real-world occupation-gender distributions in Japan, in order to shed light on how models understand Japanese culture. We believe that observing trends in these distributions over time can provide a significant lens into societal changes and transformations in gender roles within the culture.

Data. For our analysis of gender bias in Japanese language models, we first gather a list of Japanese given names, and data for computing a real-world reference distribution for Japanese occupations by gender. Japanese given names are interesting due to their sheer number of possibilities. Given names are usually composed of a sequence of one to three kanji (Chinese characters) which must be chosen from two Ministry-approved lists: the Jinmeiyō kanji (人名用漢字), a list of 863 kanji used for personal names, and the Jōyō kanji (常用漢字), a list of 2,136 commonly used characters. For this work, we scrape a list of the most popular male and female Japanese given names from the last 100 years published online by Meiji Yasuda Life Insurance Company, one of the largest insurance companies in Japan.[2] After removing duplicates, this list contains 148 female and 131 male names. We note that some of the more recent male names, such as 凪 (Nagi) and 碧 (Aoi) could also be used as female names. However, as these names are not in the female names list and hence are more commonly used as male names, we evaluate them as male names.

We also download statistics for occupations by gender from the Statistics Bureau of Japan.[3] Most of this data exists as low-quality scanned PDFs, so we perform a combination of Optical Character Recognition using Tesseract and manual transcription to extract the occupation names and counts of occupations per gender. We then remove duplicate occupations and occupations categorized as "Other", resulting in a total of 257 occupations.

[2] https://www.meijiyasuda.co.jp/enjoy/ranking/index.html

[3] www.e-stat.go.jp

Occupation	Male%	Female%
Railways line construction workers	100	0
Ships' captains, navigation officers	99.7	0.3
Carpenters	98.5	1.5
Private tutors	52.3	47.7
Artists, designers, photographers	52.1	47.9
Teachers	47.8	52.2
Nutritionists	4.6	95.4
Childcare workers	3.1	96.9
Midwives	0	100

Table 1: A selection of occupations from the 2020 census (Statistics Bureau of Japan) and the gender distributions by occupation. The occupations presented here are either dominated (i.e. $\geq 95\%$) by one gender, or have a more balanced distribution.

This data is used to compute the real-world distribution of occupations by gender.

Table 1 shows some examples of occupations dominated (i.e. $\geq 95\%$) by either gender and those that have a more balanced distribution. We find some occupations that fit traditional gender stereotypes, such as most midwives are women and most ship captains are males. On the other hand, the distribution of genders is more balanced for occupations like private tutors, artists, designers, and teachers.

Prompts. Prompts, also known as templates, have been developed to probe language models' biases (Solaiman et al., 2019; Touileb et al., 2022). We devise several Japanese prompts to probe the models' understanding of Japanese names and their associations between occupation and gender. One challenge for developing effective prompts is the Japanese SOV word order. Existing work has largely focused on SVO language like English, where an autoregressive model can take a prompt such as "[NAME] works as" and generate an occupation directly after the prompt. However, in Japanese, the phrase "works as" (として働く, toshite hataraku) comes after the occupation. Thus, to have a complete sentence, we craft the prompt [NAME]は[MASK]として働いています ([NAME] wa [MASK] toshite hataraite imasu) "[NAME] works as [MASK]"; this type of prompt is suitable for masked language models. For autoregressive language models, we first state a question of what [NAME] is working as, and then end the prompt with [NAME]は (は *wa* is a topic marker), which prompts the model to complete the sentence in the present progressive tense. This allows for a fair comparison of similar prompts for both masked and autoregressive LMs. After designing and conducting preliminary experiments with many different prompts, we select three prompts for each language model type, listed in Table 2. The prompts are designed so that they are semantically similar but differ syntactically, and can probe the occupations directly.

Models. To identify how the choice of training data as well as model training objective affects the generation of biased language, we compare several recent language models on various dimensions including English, multilingual, and Japanese language models, and masked vs. autoregressive language models. We experiment with a variety of models listed in Table 3. We include English models because previous work has shown that models trained on English data are able to produce multilingual generations (Radford et al., 2019). Note that the Japanese models are trained from scratch on Japanese text, not fine-tuned existing models. For models available from HuggingFace, we experiment with greedy decoding and constrained decoding. Constrained decoding is a method in which we can guide the generation to a specific output. In our case, we constrain the model to output occupations by forcing the model (using the `force_words_ids` parameter in HuggingFace) to output specific tokens when it encounters the [MASK] token (for masked models) or is at the end of the prompt (for autoregressive models), and we calculate the probability of that specific occupation. When the occupation is represented with multiple subwords, the prediction is performed based on the full occupation name. We repeat this process for every occupation to create a probability distribution over all occupations for a given Japanese name. For Bard and ChatGPT, we take the output as given by the model, limited to 5 characters.

Evaluation. Using the above prompts, we generate text completions from the various models. From the generated output, we identify the occupation by extracting the first noun that appears in the generated output. We take the first noun because the model output often begins with non-noun characters, including: a comma, the unknown token ([UNK]), various (様々 samazama, さまざま samazama, and 多様 tayō), you (あなた anata), job (仕事 shigoto), unknown (不明 fumei), myriad (多彩 tasai), and same (同じ onaji).

We identify bias in these models by framing the generation as the task of gender classification, where the model is given a Japanese name and predicts gender based on the majority gender of the outputted occupation. While generating the probabilities, we consider each name separately, but at evaluation, it is performed per occupation, and not per name. Specifically, we follow Touileb et al. (2022) and report F1 scores for our models with the output generated with constrained

	#	Japanese	English Translation
Masked	1	[NAME]の職業は[MASK]です。	[NAME]'s profession is [MASK].
	2	[NAME]の職種は[MASK]です。	[NAME]'s type of occupation (job category) is [MASK].
	3	[NAME]は何の仕事をしていますか？ [NAME]は[MASK]として働いています。	What is [NAME]'s job? [NAME] works as [MASK].
Autoreg	1	[NAME]の職業は	[NAME]'s profession is
	2	[NAME]の職種は	[NAME]'s type of occupation (job category) is
	3	[NAME]は何の仕事をしていますか？ [NAME]は	What is [NAME]'s job? [NAME]

Table 2: Prompts for the masked and autoregressive language models.

Model	Language	Size
mBERT	Multilingual	110M
BERT-J	Japanese	110M
GPT-2	English	355M
mGPT	Multilingual	1.3B
GPT-2-J	Japanese	336M
GPT-NeoX-J	Japanese	3.6B
GPT-NeoX	English	20B
Bard	Multilingual	137B
ChatGPT	Multilingual	175B

Table 3: List of masked, autoregressive, and commercial LMs we experimented with.

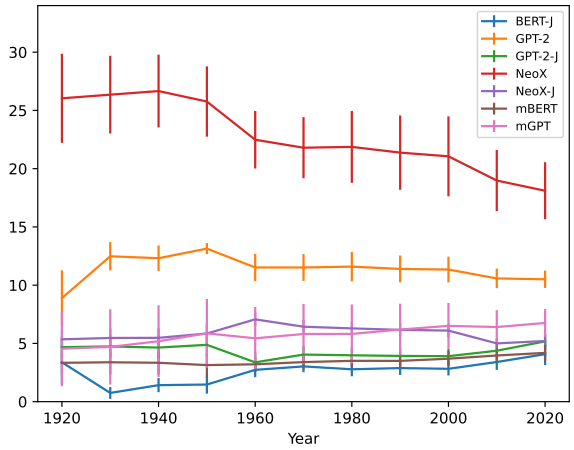

Figure 1: The mean and standard deviation of the Kullback–Leibler divergence (relative entropy) for each model for Prompt 1 ([NAME]'s profession is), compared to the census data from 1920 to 2020. Prompt 2 and Prompt 3 show similar patterns.

decoding, where the highest probability occupation is counted as correct if its majority gender from the census data matches the gender of the given name. For example, if a model outputs 大工 (daiku) *carpenter* for the male name 正一 (Shoichi), this would be a correct classification because according to the Japanese census data, more carpenters are male than female. A higher F1 score indicates that the model output aligns more with the census data.

In addition, constrained decoding allows us to compute a distribution over all occupations for a given name, and analyze this distribution against historical occupation statistics using Kullback-Leibler (KL) divergence, a measure of how one probability distribution is different from another distribution. We compare all models against the 2020 census data, then conduct temporal analyses by comparing the KL divergence values across each decade from 1920 to 2020, then compare the values across each decade for names from different decades to study how models are representative of occupational biases from different time periods.

4. Results and Analyses

4.1. Constrained Decoding

We first evaluate the gender classification experiments with constrained decoding on all occupations combined, where the majority gender (according to the census data) of the occupation with the highest model probability is taken to be the prediction. Results broken down by gender of the name are summarized in Table 4. Out of the models we tested, BERT-Japanese achieves the highest performance, which we believe is because it was trained specifically on Japanese text, and also because its masked language modeling is well-suited to the SOV word order of Japanese. We find that Prompt 1 seems to be most effective for masked language models, while Prompt 3 is more effective for autoregressive models. We analyze the specific prompts in the following sections and then follow with a temporal analysis comparing against the real-world distribution of occupations.

Prompt	mBERT			BERT-J			GPT-2			mGPT			GPT-2-J			GPT-NeoX-J			GPT-NeoX		
	F	M	T	F	M	T	F	M	T	F	M	T	F	M	T	F	M	T	F	M	T
1	.31	**.79**	.54	.37	**.87**	.61	**.61**	.46	.54	**.50**	.33	.42	**.68**	.13	.42	.29	**.72**	.49	.12	**.76**	.42
2	.42	**.56**	.49	**.94**	.12	.56	**.68**	.40	.55	**.57**	.54	.56	.27	**.78**	.51	.38	**.56**	.46	.49	**.50**	.49
3	.33	**.53**	.42	**.79**	.23	.53	.23	**.58**	.46	.27	**.66**	.45	.61	**.66**	.64	.16	**.87**	.49	**.60**	.41	.51

Table 4: F1 scores of models on the three prompts, compared to the actual occupation distribution for all 257 occupations. Overall, BERT-Japanese performed the best, due to being trained on Japanese text and the masked language modeling matching the SOV order of Japanese. F, M, and T indicate the set of names used: female, male, and a combined set (total, through micro-averaging), respectively.

Prompt	mBERT	BERT-J	GPT-2	mGPT	GPT-2-J	GPT-NeoX-J	GPT-NeoX
1	4.39 ± 0.599	4.54 ± 0.939	10.6 ± 0.670	6.91 ± 1.44	5.31 ± 0.643	5.40 ± 0.181	17.8 ± 2.20
2	4.42 ± 0.628	4.12 ± 0.722	10.3 ± 0.802	7.14 ± 1.89	4.48 ± 0.543	5.28 ± 0.110	19.7 ± 1.76
3	3.13 ± 0.328	5.68 ± 0.916	9.86 ± 0.137	4.93 ± 0.387	4.83 ± 0.640	5.32 ± 0.230	15.2 ± 0.688

Table 5: The mean and standard deviation of KL divergence between the model distribution and 2020 census data, averaged over all names. A smaller value indicates that the model's distribution is more similar to the census data; mBERT, GPT-2-Japanese, and GPT-NeoX-Japanese had notably small divergences.

Prompt 1 (profession). The first prompt asks for [NAME]'s 職業 (shokugyō), which translates to profession or occupation. As shown in Table 4, BERT-Japanese and GPT-2-Japanese perform best overall. For the first prompt, mBERT, BERT-Japanese, GPT-NeoX-Japanese, and GPT-NeoX achieve high F1 scores for the male names with substantially lower scores for the female names, with BERT-Japanese achieving the highest score of 0.87. The remaining three models perform better for female names, with GPT-2-Japanese achieving the highest score of 0.68. Most models perform significantly better on one gender than the other, though which gender depends on the model, in contrast to Touileb et al. (2022) who found that models consistently performed better on male-leaning Norwegian occupations.

Examining the KL divergence between the distribution of predicted occupations and the real-world gender distribution in Table 5, we find that BERT-Japanese shows the smallest divergence, followed by mBERT and GPT-2-Japanese. Again, we believe that the combination of training on Japanese data, as well as the structure of the task better suiting the masked language models, allows the models to better capture the real-world distribution of occupations and genders.

Figure 2 displays the normalized maximum probabilities for each occupation category found in the 2020 census data for Prompt 1. Most notable are the probabilities for professional and technical occupations, 0.781 and 0.832 (compared to the real-world distribution is 0.199 and 0.170), for female names and male names respectively, for GPT-2-Japanese. This is due to the high correlation between *nurse* and *private tutor* for female names and *writer/editor* and *judge/lawyer* for male names. We also see a high probability for haulage/packaging/cleaning-related occupations in both genders using GPT-NeoX-Japanese. Our results indicate that these two Japanese models have a preference for a small set of occupations, resulting in skewed probabilities.

Prompt 2 (job category). The second prompt asks for the [NAME]'s 職種 (shokushu), the type of occupation or job category. BERT-Japanese and GPT-2-Japanese are the best performing for this prompt as well, although for different genders. Except for mGPT and GPT-NeoX, the models are again skewed toward one gender. However, BERT-Japanese and GPT-2-Japanese switch their high-scoring classes, with BERT-Japanese achieving 0.94 for female names and GPT-2-Japanese achieving 0.78 for male names. These results suggest that the two Japanese models may not be skewed toward either gender but rather have a tendency to output a certain occupation depending on the wording of the prompt. Again, the KL divergence with the real-world distribution is much lower for mBERT, BERT-Japanese, and GPT-2-Japanese, which we believe is due to similar reasons as for Prompt 1.

Examining the normalized probabilities by occupation category, most distributions are similar for both genders, except for transportation/mechanical-related occupations for BERT-Japanese, which shows a high percentage (.432) for female names compared to male names (.239). This does not correlate with the real-world distributions of each gender nor with the total distribution across both genders. Analyzing the individual probabilities for occupations within this category,

Names	mBERT	BERT-J	GPT-2	mGPT	GPT-2-J
正一 (Shoichi)	father*	carpenter	-	teacher	police
豊 (Yutaka)	Yutaka*	carpenter	-	medical technologist	doctor
歩夢 (Ayumu)	voice*	singer	human*	caregiver	detective
キヨ (Kiyo)	voice*	housewife	-	designer	-
千尋 (Chihiro)	voice*	secretary	-	pirate*	Chihiro*
芽依 (Mei)	girl*	model	fire*	nurse	nurse
Names	GPT-NeoX-J	GPT-NeoX	Bard	ChatGPT	
正一 (Shoichi)	director	teacher	construction	doctor	
豊 (Yutaka)	bartender	handmade*	farmer	farmer	
歩夢 (Ayumu)	singer	wizard*	singer	actor	
キヨ (Kiyo)	Kiyo*	handmade*	nursery teacher	florist	
千尋 (Chihiro)	scientist	wizard*	translator	lawyer	
芽依 (Mei)	nurse	food processing	nurse	paralegal	

Table 6: Example output sampled from the models. Shoichi, Yutaka, and Ayumu are common male names, and Kiyo, Chihiro, and Mei are common female names. A hyphen - indicates either that the generation was not in Japanese or that the generated text consisted of random characters. An asterisk * indicates Japanese text was generated but was not an occupation. The Japanese models as well as Bard and ChatGPT tended to produce better output.

such as *captain/navigator* or *automobile driver*, male names are actually more likely to be associated with these occupations than female names. However, other occupation categories also have high probability for male names, so normalizing across occupation categories results in a flatter distribution compared to female names.

Prompt 3 (works as). Prompt 3 asks what [NAME] is doing (している shite iru) as their job. BERT-Japanese and GPT-2-Japanese are still strong-performing models. However, for male names, GPT-NeoX-Japanese has the highest F1 score of 0.87, which is the highest of all scores for this prompt. This suggests that this longer prompt helps boost the performance of GPT-NeoX-Japanese, while showing a small decline for BERT-Japanese and GPT-2-Japanese. Adding a complete sentence within the prompt may have helped guide the larger autoregressive model to generate an occupation, instead of a different continuation of the prompt.

Analyzing the KL divergence for this prompt, we observe that mBERT, GPT-2-Japanese, and GPT-NeoX-Japanese still most closely match real-world distributions, but are now followed by mGPT. This suggests that the change in prompt wording also affects the generation with multilingual models as well. When focusing on the performance of GPT-NeoX-Japanese, the F1 score suggests a possible improvement, while KL divergence suggests minimal change. We note that for all multilingual and English models, Prompt 3 had a lower KL divergence. However, for the three Japanese models, it did not improve the divergence and for BERT-Japanese had a slight decline. Thus, the structure of this prompt may have indicated to non-Japanese models that an occupation is to be generated, while for Japanese models, this did not provide any new information over Prompts 1 and 2.

We observe that for BERT-Japanese, GPT-2-Japanese, and GPT-NeoX-Japanese, high-probability occupation categories for Prompts 1 and 2 show similar trends with Prompt 3. However, we see that unlike Prompts 1 and 2, here mGPT heavily favors professional and technical occupations: 0.798 and 0.832 for female and male names respectively, which are the highest across all models and all categories. In addition to the occupations covered by GPT-2-Japanese, mGPT predicts high probabilities for occupations such as *accountant*, *therapist*, and *social worker*. The high probabilities are consistent across the two genders.

4.2. Bard and ChatGPT

As we do not have local access to these models, we do not have probabilities associated with their output. Thus, we conduct a more qualitative analysis of these models' output. Overall, these models generate fluent output, often much lengthier than the other models we experimented with. They can also handle both the masked and autoregressive prompts.

Bard. Given the masked LM prompts and limited to five-character generations, Bard sometimes explains why it generated a certain output. For Prompt 2 (type of occupation), it explained how the

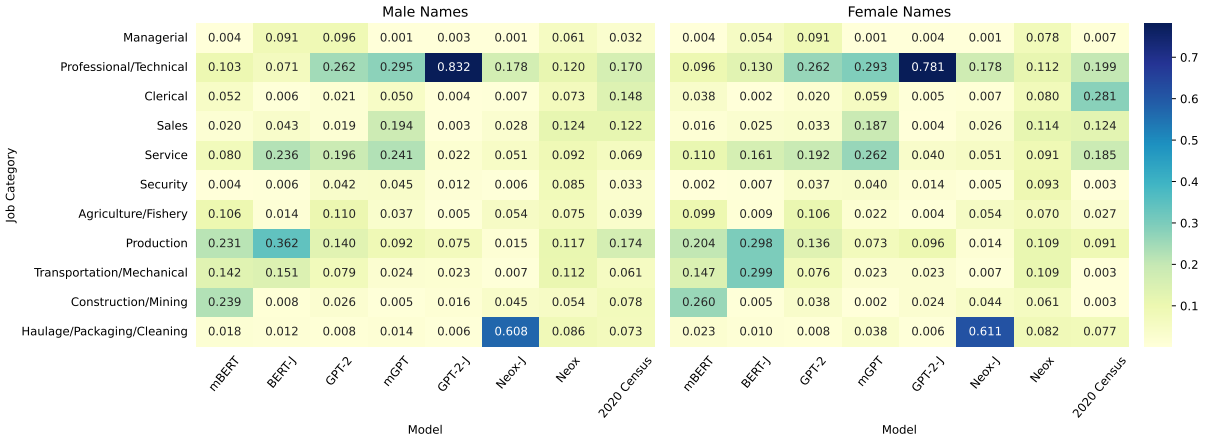

Figure 2: Normalized maximum probabilities of an occupation in each occupation category, provided in the 2020 census data, averaged over all male names and all female names for each model. As the probabilities were generated per name, and evaluated per occupation, they do not sum to one across the occupations, and thus required normalization across the occupations. The rightmost column in each heatmap is the real-world distribution of occupations for each gender, from the 2020 census. This distribution is for Prompt 1.

generated occupations were associated with the names' meanings or sounds. For example, Bard noted that the name 蒼空 (Sora) contained characters related to the sky, and thus suggested that Sora's occupation was a *pilot* or *astronaut*.

Bard did not give any explanations for Prompts 1 and 3, but in some cases, we found that Bard may have associated the name with an occupation based on some real-world knowledge. For example, Bard associated the name 翔平 (Shohei) with *professional baseball player*, likely due to Shohei Ohtani, and the name 博之 (Hiroyuki) with *businessman/entrepreneur*, likely due to Hiroyuki Nishimura, an internet entrepreneur. These names exist in Wikipedia, which Bard was trained on, and we see that Bard is able to make use of this knowledge.

ChatGPT. For some prompts, ChatGPT returned that it was unable to determine the [NAME]'s occupation and asked for additional information. In other cases, it replied that it was unable to give a definitive or deterministic answer but still provided possible occupations while noting that these were typical occupations and not specific to the given name. This occurred for all three prompts.

Manually examining the generated text, we were not able to identify any unique occupations related to public figures, but we saw some connections between the names' characters and occupations. For example, for the name Takumi (拓海), which uses the character for sea (海), ChatGPT generated occupation *fisherman*. Thus, ChatGPT may (erroneously) use its knowledge of the meaning of characters in the name to output an occupation.

4.3. Temporal Analysis

Names fluctuate in popularity over time. In this section, we examine whether the associations between occupation and gender in language models are more representative of a certain time period. Comparing the seven models against the occupation data of the last 100 years, we compute the average KL divergence between names from each decade and the real-world distribution from that decade. We did not observe a significant difference in the KL divergence across occupation statistics from given years, or for different prompts as shown in Figure 1. The KL divergence for GPT-NeoX and GPT-2 was the largest throughout the entire century, with GPT-NeoX showing a slight decrease throughout time. For the other five models, the KL divergence scores were relatively stable, suggesting that the trends seen in the generated text were not representative of a certain time period but rather constant throughout.

When examining the Japanese models, BERT-Japanese, which is trained on Japanese Wikipedia, shows the lowest KL divergence, compared to GPT-2-Japanese, trained on Japanese CC-100 and Japanese Wikipedia, and NeoX-Japanese, trained on Japanese Wikipedia, Japanese CC-100, and Japanese C4. This suggests that the occupation and gender associations in text from Japanese Wikipedia most closely match the real-world data distribution. On the other hand, the Japanese models consistently showed lower KL divergence than the English and multilingual models, indicating that these models

may have picked up gender and occupation associations from other cultures that differ from those of Japan.

To examine the association between names and time period, we also compared the KL divergence for subsets of names, grouping the names by every two decades, shown in Figure 3. Each curve represents the divergence for the names that were popular within a certain two-decade time period. If the model changes its distribution of outputted occupations based on when the name was popular, then we should see five different curves with minima at each vicennial. For example, for the curve representing 1920 to 1940, if the models had an accurate association of names at the time and occupations at the time, the divergence for the curve would be the smallest in the 1920-1940 range. We do not see this phenomenon, and thus conclude that these models associate gender and occupation regardless of the time period.

4.4. Comparison with Greedy Decoding

Initially, we experimented with using greedy decoding to obtain the most likely text after the topic marker は in the prompts. After extracting the occupations from the generated text, if it exists, we compare the edit distance with all of the occupations from the census data and deem them as a match only when the edit distance is zero. Unlike languages that use the alphabet, in Japanese, one character can encompass a lot of information and in some cases, the occupation titles are three characters long; thus, allowing an edit distance of one as a match would yield incorrect counts for each occupation.

Using greedy decoding did not work as intended, due to two reasons. First, the multilingual models had difficulty producing occupation titles, while Japanese models, especially BERT-Japanese, were able to generate them, albeit with a small set of occupations. The multilingual models often generated characters such as 男 (otoko) *man* and 女 (onna) *woman*, the name of the person in the prompt, or something that was not an occupation. For the Japanese models, although the masked language models generally produce the expected occupations, the autoregressive GPT-NeoX-Japanese model produces text that continues the prompt but does not include an occupation. Second, the Japanese language models are able to generate complete sentences or tokens, but in most cases, still fail to generate an occupation that matches the census data. The range of occupations that did match was limited: for male names, almost all generated occupations were *police officer*, *detective*, or *carpenter*, and for female names, almost all were *hairdresser* or *nurse*. This reinforces our findings that models tend to prefer a small range of occupations. Table 6 contains a sample of occupations generated by the models. Because of these challenges, and to understand the entire distribution over occupations, we evaluated the models' output using constrained decoding to force models to output all occupations.

5. Conclusion

We have investigated how gender and occupations are represented in pre-trained language models, and how these distributions correlate with real-world statistics, focusing on how models associate Japanese names with gender and occupations. From our three prompts combining names and occupations tailored to both masked and autoregressive LMs, we find that models trained solely on Japanese text generate distributions of occupations that are closer to the real-world distribution than non-Japanese models. With some prompts, multilingual models can achieve comparable results. Although models can perform well with constrained decoding, when performing sampling or greedy decoding, multilingual models (that have seen Japanese) are not able to understand the prompt and generate occupations, and Japanese-specific models resort to a limited set of stereotypically gendered occupations. These findings highlight the need for researchers to take these biases into account in downstream tasks, even for high-resource languages like Japanese. In future work, we plan to investigate occupational distributions from other countries and examine the extent to which language-specific language models reflect real-world distributions.

Ethics Statement

It is critical to study the potential biases in language models so that biases are not amplified in downstream tasks (Bender et al., 2021). We expect that language models include bias from their training corpora, and it is a fact that not all occupational distributions are evenly split between female and male workers. However, we must acknowledge that disparities exist in real-world distributions, and that this can be reflected in language models. Our work is uncovering such disparities, and we urge researchers to design future experiments and applications carefully with this in mind, so as to not amplify any bias included.

Limitations

A limitation of our work is that we are only able to evaluate the correlation between the binary gender categories of female/male with occupations due to the availability of statistics for these genders. We

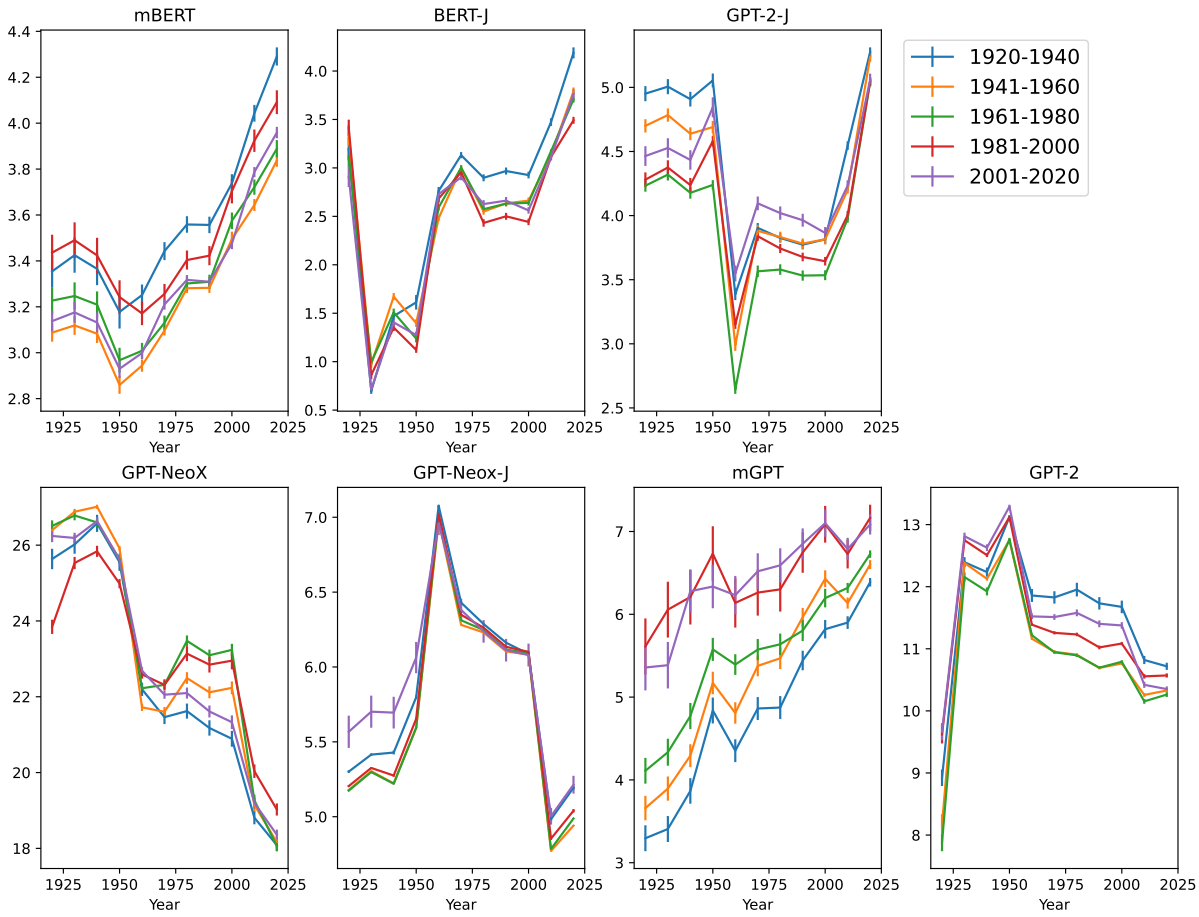

Figure 3: The mean and standard error of the Kullback–Leibler divergence (relative entropy) for all seven models for Prompt 1, compared to the census data from 1920 to 2020. The horizontal axis labels indicate the census data, while the line colors represent the subset of names popular within a certain two-decade period. The plot shows that the period when a certain name was common did not correlate with the occupational distribution at that period.

acknowledge that gender includes a wider spectrum than this. Furthermore, we are not able to test with the latest models such as GPT-4. However, ethical restrictions built into these models, such as we have seen with ChatGPT, may prevent some of this analysis.

Additionally, we acknowledge that gender bias is not the only type of bias that language models may exhibit. However, considering the nuances and intricacies inherent in the Japanese language and the lack of in-depth research exploring this particular area, a focused investigation was necessary. In the future, we aim to extend our research to study other forms of bias, such as racial or socio-economic bias, building on the foundation provided by our current work.

Concerning the ethical implications of bias in language models, our intention was to shed light on the existence of bias in Japanese language models and kick-start the conversation within the scientific community. An in-depth discussion of specific ethical concerns and potential solutions would have added critical insight, but such analysis would easily expand beyond the scope of our current paper and require expert input from fields like ethics and sociology. As we continue and expand our research, we plan to seek interdisciplinary collaborations to engage these important topics more thoroughly.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable suggestions.

References

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

gies, pages 2824–2830, Seattle, United States. 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*, pages 610–623.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

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

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

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Youki Itoh and Hiroyuki Shinnou. 2021. Domain-specific Japanese ELECTRA model using a small corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 640–646, Held Online. INCOMA Ltd.

Charles Li and Sandra Thompson. 1976. Subject and topic: A new typology of language. *Subject and Topic*, pages 457–489.

Keisuke Miyazaki, Hiroaki Yamada, and Takenobu Tokunaga. 2022. Cross-domain analysis on Japanese legal pretrained language models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 274–281, Online only. Association for Computational Linguistics.

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.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

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

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.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, and Takashi Ninomiya. 2022. A Japanese masked language model for academic domain. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 152–157, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Appendix

Appendix A. Language Models Used

Table 7 is the complete version of Table 1, with the full model names provided on Hugging Face, along with how we referred to the models within the paper. For the masked and autoregressive models, links are provided to the respective Hugging Face pages.

Below are the training data sources for all seven masked and autoregressive models.

- *bert-base-multilingual-cased* is trained on 104 languages with the largest Wikipedias; the list of languages can be found here
- *bert-base-japanese* is trained on Japanese Wikipedia
- *gpt2-medium* is trained on WebText, a dataset created by scraping all web pages from outbound links on Reddit which received at least 3 karma
- *mGPT* is trained on 61 languages from 25 language families using Wikipedia and Colossal Clean Crawled Corpus (C4)
- *japanese-gpt-medium* is trained on Japanese CC-100 and Japanese Wikipedia
- *japanese-gpt-neox-3.6b* is trained on Japanese CC-100, Japanese C4, and Japanese Wikipedia
- *GPT-NeoX* is trained on the Pile

Model	Name Used in Paper	Language	Size
google-bert/bert-base-multilingual-cased	mBERT	Multilingual	110M
tohoku-nlp/bert-base-japanese	BERT-J	Japanese	110M
openai-community/gpt2-medium	GPT-2	English	355M
ai-forever/mGPT	mGPT	Multilingual	1.3B
rinna/japanese-gpt2-medium	GPT-2-J	Japanese	336M
rinna/japanese-gpt-neox-3.6b	GPT-NeoX-J	Japanese	3.6B
GPT-NeoX	GPT-NeoX	English	20B
Bard	Bard	Multilingual	137B
ChatGPT	ChatGPT	Multilingual	175B

Table 7: List of masked, autoregressive, and commercial LMs we experimented with.

Appendix B. Kullback-Leibler Divergence for All Prompts

Below are the KL divergence for each model, for all three prompts.

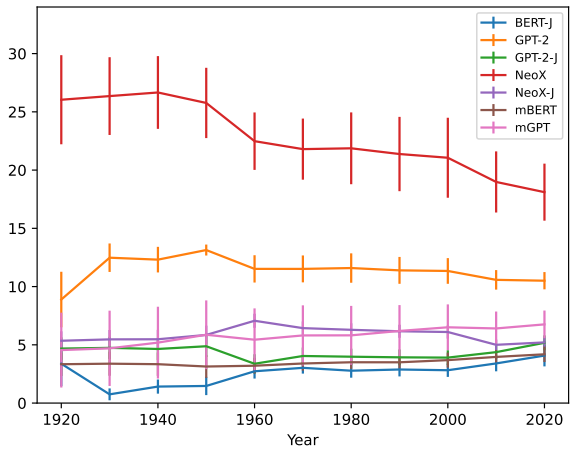

Figure 4: The mean and standard deviation of the Kullback–Leibler divergence (relative entropy) for each model for Prompt 1 ([NAME]'s profession is), compared to the census data from 1920 to 2020.

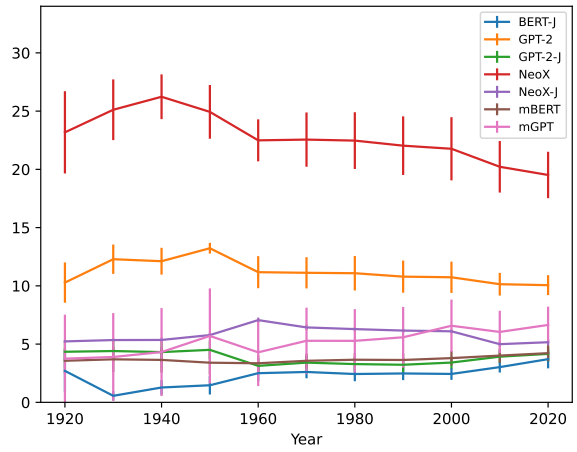

Figure 5: The mean and standard deviation of the Kullback–Leibler divergence (relative entropy) for each model for Prompt 2 ([NAME]'s type of occupation (job category) is), compared to the census data from 1920 to 2020.

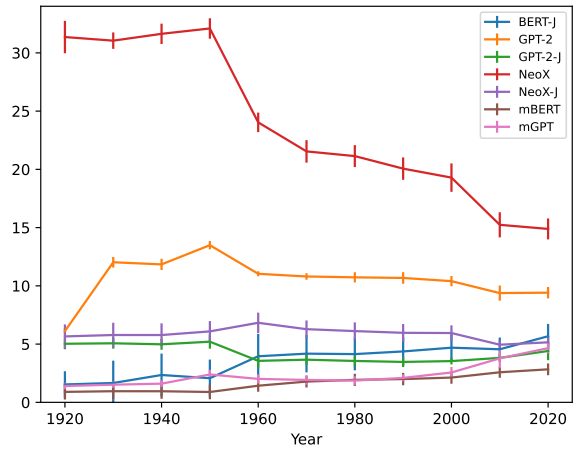

Figure 6: The mean and standard deviation of the Kullback–Leibler divergence (relative entropy) for each model for Prompt 3 (What is [NAME]'s job? [NAME] works as), compared to the census data from 1920 to 2020.

Appendix C. Normalized Maximum Probabilities for All Prompts

Below are the normalized maximum probabilities of an occupation, for all three prompts.

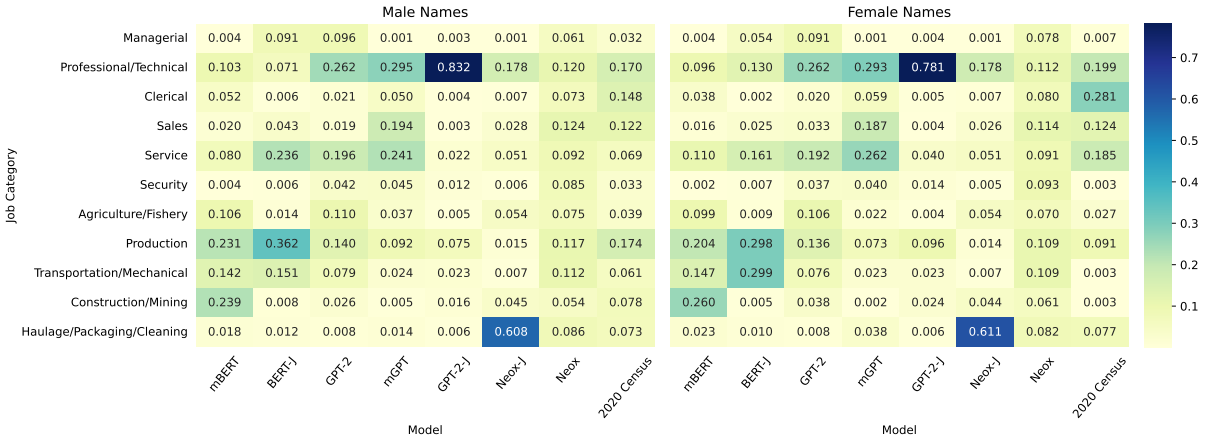

Figure 7: Normalized maximum probabilities of an occupation in each occupation category, provided in the 2020 census data, averaged over all male names and all female names for each model. As the probabilities were generated per name, and evaluated per occupation, they do not sum to one across the occupations, and thus required normalization across the occupations. The rightmost column in each heatmap is the real-world distribution of occupations for each gender, from the 2020 census. This distribution is for Prompt 1 ([NAME]'s profession is).

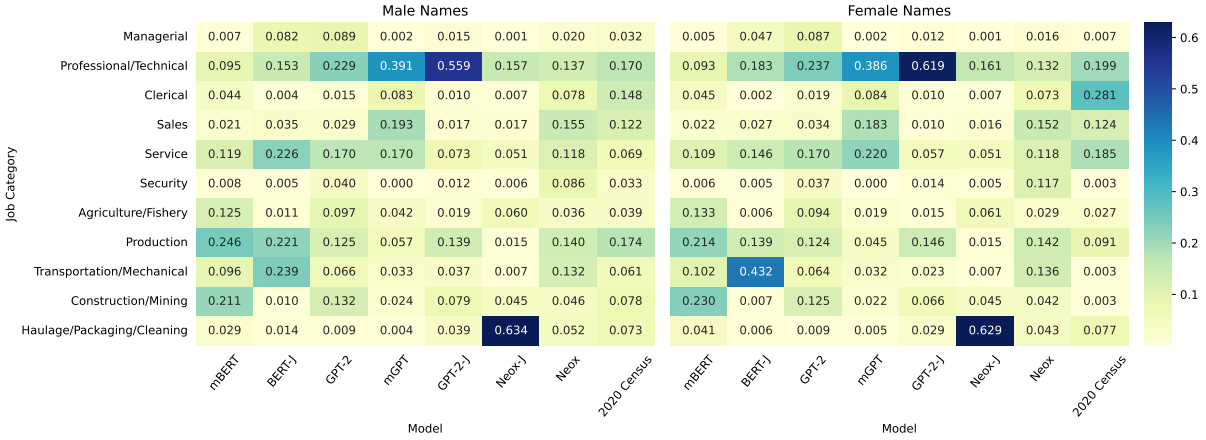

Figure 8: Normalized maximum probabilities of an occupation in each occupation category, provided in the 2020 census data, averaged over all male names and all female names for each model. As the probabilities were generated per name, and evaluated per occupation, they do not sum to one across the occupations, and thus required normalization across the occupations. The rightmost column in each heatmap is the real-world distribution of occupations for each gender, from the 2020 census. This distribution is for Prompt 2 ([NAME]'s type of occupation (job category) is).

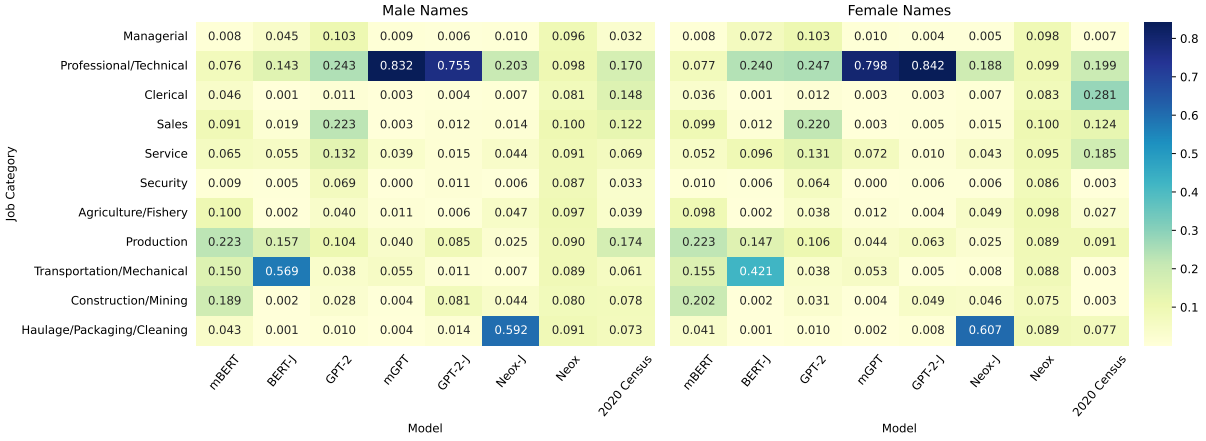

Figure 9: Normalized maximum probabilities of an occupation in each occupation category, provided in the 2020 census data, averaged over all male names and all female names for each model. As the probabilities were generated per name, and evaluated per occupation, they do not sum to one across the occupations, and thus required normalization across the occupations. The rightmost column in each heatmap is the real-world distribution of occupations for each gender, from the 2020 census. This distribution is for Prompt 3 (What is [NAME]'s job? [NAME] works as).