Examining the Causal Effect of First Names on Language Models: The Case of Social Commonsense Reasoning

Sullam Jeoung   Jana Diesner   Halil Kilicoglu
University of Illinois at Urbana-Champaign
{sjeoung2, jdiesner ,halil}@illinois.edu

Abstract

As language models continue to be integrated into applications of personal and societal relevance, ensuring these models' trustworthiness is crucial, particularly with respect to producing consistent outputs regardless of sensitive attributes. Given that first names may serve as proxies for (intersectional) socio-demographic representations, it is imperative to examine the impact of first names on commonsense reasoning capabilities. In this paper, we study whether a model's reasoning given a specific input differs based on the first names provided. Our underlying assumption is that the reasoning about *Alice* should not differ from the reasoning about *James*. We propose and implement a controlled experimental framework to measure the causal effect of first names on commonsense reasoning, enabling us to distinguish between model predictions due to chance and caused by actual factors of interest. Our results indicate that the frequency of first names has a direct effect on model prediction, with less frequent names yielding divergent predictions compared to more frequent names. To gain insights into the internal mechanisms of models that are contributing to these behaviors, we also conduct an in-depth explainable analysis. Overall, our findings suggest that to ensure model robustness, it is essential to augment datasets with more diverse first names during the configuration stage.

1 Introduction

Recent language models (LMs) (Brown et al., 2020; Radford et al., 2019) have shown remarkable improvements when used in NLP tasks and are increasingly used across various application domains to engage with users and address their personal and social needs, such as AI-assisted autocomplete and counseling (Hovy and Yang, 2021; Sharma et al., 2021). As these LMs models are adopted, their social intelligence and commonsense reasoning have become more important, especially as AI models
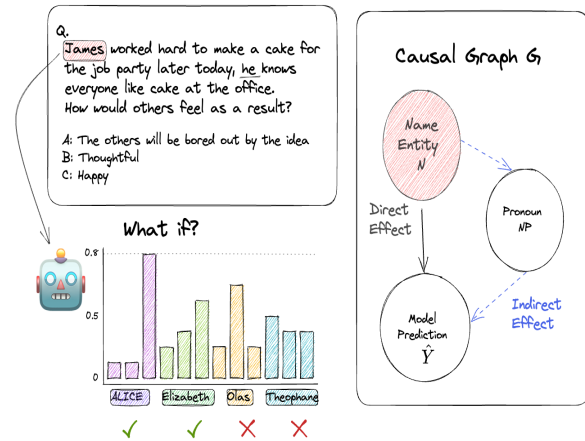

Figure 1: Framework of our approach. (Left): An example template with name instances (Right): The causal graph G we hypothesize for analysis

are deployed in situations requiring social skills (Wang et al., 2007, 2019).

In this paper, we examine how first names are handled in commonsense reasoning (Fig 1). To this end, we measure the causal effect that name instances have on LMs' commonsense reasoning abilities. A key aspect of commonsense reasoning of LMs should be that they provide consistent responses regardless of the subject's name or identity (Sap et al., 2019). That is, the reasoning behind "*Alice*" should not differ from that about "*James*", for instance. Given that first names can be a proxy for representation of gender and/ or race, this consistency is essential not only for the robustness but also for the fairness and utility of a LM.

Previous studies have revealed that pre-trained language models are susceptible to biases related to peoples' first names. For instance, in the context of sentiment analysis, certain names have been consistently associated with negative sentiments by language models (Prabhakaran et al., 2019). Additionally, during text generation, names have been found to be linked to well-known public figures, indicating biased representations of names

(Shwartz et al., 2020). Furthermore, Wolfe and Caliskan (2021) demonstrated that less common names are more likely to be 'subtokenized' and associated with negative sentiments compared to frequent names. These studies shed light on how pre-trained language models disproportionately process name representations, potentially leading to biased outputs.

While examining pre-trained language models is valuable to understand their capabilities and limitations, in many cases the models are fine-tuned, or adapted and optimized, to guarantee improved performance on specific downstream tasks, such as text classification, machine translation, and question answering, among others (Bai et al., 2004; Peng and Dean, 2007; Rajpurkar et al., 2018). Given that fine-tuning pre-trained language models can lead to major performance gains (Devlin et al., 2019), in this paper, we ask if performance disparities based on names still exist even when the models are fine-tuned. If so, we ask which components of the models contribute to performance disparities and to what extent. We design a controlled experimental setting to determine whether performance differences arise by chance or are caused by names. Our contributions are three-fold[1]:

- We propose a controlled experimental framework based on a causal graph to discern the causal effect of first names in the commonsense reasoning of language models. We leverage the name statistics from U.S. Census data for this purpose.
- We present an in-depth analysis to understand the internal model mechanisms in processing first names. To be specific, we examine the embeddings and neuron activation of first names.
- Based on our analysis, we provide suggestions for researchers in configuring the datasets to provide more robust language modeling.

2 Task Formulation

We consider a dataset of commonsense reasoning examples $d \in \mathcal{D}$, where each item consists of a question $q \in \mathcal{Q}$, three possible answer candidates $\mathcal{C} = \{c_1, c_2, c_3\}$, and a label $y \in Y$, which is the correct answer among the candidates. $\mathcal{Q}$ and $\mathcal{C}$ serve as a template t, containing placeholders for names $[n]$ and pronouns referring to the names,

$[np]$.

To ensure grammatical correctness, a pronoun placeholder np is set in variants of subject pronoun np_1, object pronoun np_2, and dependent possessive pronouns np_3. An example of the data template is as follows:

Question $\mathcal{Q}$: Typically every four months, [n] went to the doctor for a routine checkup and was told $[np_1]$ needs rest. What will [n] want to do next?

Candidates $\mathcal{C}$:{
(a) call the doctor, (b) finish all $[np_2]$ projects and postpone the rest, (c) take time off from work}
Label y: (c) take time off from work

---

3 Causal Graph

A language model can be denoted as a function f, taking inputs as follows:

$$\hat{y} = f(t(\boldsymbol{n}, \boldsymbol{np})) \tag{1}$$

We are interested in how first names ($\boldsymbol{n} \in N$) influence the prediction $\hat{y} \in \hat{Y}$ under the function f. We hypothesize that there is a causal graph $\mathcal{G}$ that encodes possible causal paths relating first names to the model's prediction (Fig 1, right). [2]
We identify both the direct effect and indirect effect on model prediction (Pearl, 2022):

1. The direct effect of names on model prediction ($N \rightarrow \hat{Y}$) measures how names have a direct impact on model predictions (without going through any intermediate variables).

2. The indirect effect indicates potential confounding factors associated with names that may influence predictions. We hypothesize that pronouns are an intermediate variable ($N \rightarrow NP \rightarrow \hat{Y}$). Intuitively, pronouns that refer to names can influence how models make their predictions. For example, this indirect effect indicates changes in model prediction when the pronouns differ (e.g. *he* vs. *she*) but the names remain the same or fixed (e.g. *Pat*). Pronouns inherently associate with the names they refer to, and this association may cue models to consider those names more strongly when generating a response. Thus, we posit the effect of pronouns as an indirect effect.

---

[2]Specifically, when referring to the causal graph, it pertains to the utilization of causal directed-acyclic graphs (DAGs), as mentioned in the work by (Feder et al., 2022)

Below, we formalize the causal mechanisms, intervention lists, and the effect size that measures the change in model prediction.

Direct Effect

$$\text{DE}(N \to \hat{Y}) :=$$
$$\sum_t \mathbb{E}_N^+[\hat{Y}|T = t] - \mathbb{E}_N^-[\hat{Y}|T = t]$$

where $\mathbb{E}_N^+[\hat{Y}|T = t]$ indicates the average effect size of name lists N^+, while $\mathbb{E}_N^-[\hat{Y}|T = t]$ indicates the average effect size of name lists N^- on template t. The details of the name lists of interest N^+ and N^- are listed in section 3.1 and the effect size is defined in section 3.2. DE measures the causal effects between name lists via direct do-interventions of N^+ as the template t is fixed (Pearl, 1995). Beyond computing the differences, to test the null hypothesis, we conduct a t-test and obtain the p-value statistics.

Indirect effect

$$\text{IE}(N \to \hat{Y}) := \sum_t^T \sum_n^N (\mathbb{E}_{NP}^+[\hat{Y}|T = t, N = n]$$
$$- \mathbb{E}_{NP}^-[\hat{Y}|T = t, N = n])$$

where $\mathbb{E}_{NP}^+[\hat{Y}|T = t, N = n]$ indicates the average prediction conditioned on template t and name n, with the set of NP^+, and $\mathbb{E}_{NP}^-[\hat{Y}|T = t, N = n]$ refers that of NP^-. To account for the effect of names, note that names are also controlled along with the template.

3.1 Causal Intervention

We apply feasible intervention on T : $\{q, c, (n, np), y\}$ to T' : $\{q, c, (n', np'), y\}$. We denote the intervention list as $\text{Do}(X : x \to x')$, where $X \in \{Q, C, (N, NP), Y\}$. We denote $\hat{y}' \in \hat{Y}'$ to indicate the prediction of the intervened X'. As we want to explore names based on their characteristics, we partition the intervention lists N based on two criteria: *frequency* and *gender*. These criteria were chosen following previous work (Wolfe and Caliskan, 2021; Buolamwini and Gebru, 2018) that has demonstrated that less common names, as well as gender, can be key factors in models that exhibit biases. Studies have shown that models trained on datasets with an imbalance of names or gender can reflect and even amplify prejudices, resulting in unfair outcomes, particularly for marginalized groups (Bolukbasi et al., 2016; Zhao et al., 2017). By focusing on name frequency and gender representation, we aim to evaluate the impact of these criteria on models. In order to base our work on prior statistics, we use the name statistics from the U.S. Census data. The detailed process of how the intervention list was filtered from the dataset is outlined in section 5. We consider the set of names for do-intervention as below:

MOST-LEAST Based on the frequency of names, N_{MOST} indicates the names with top-k highest frequency, whereas N_{LEAST} refers to lowest frequency.

FEMALE-MALE We use the gender information from the statistics to discern the gender of a name. Note that we purely refer to the 'gender' of names based on their records. That is, we account for cases where a name can be both male or female, based on the frequency statistics. For example, if the records for *Lee* exist for both males and females, we consider the name belonging to both genders to reflect real-world data.

3.2 Effect Size

To evaluate the impact of our model, we utilize two distinct metrics.

ACCURACY To quantify the degree of wrong predictions, we define $\mathbf{d}_{\text{ACC}}$ as

$$\mathbf{d}_{\text{ACC}}(x) := \mathbb{1}(\hat{y} \neq y)$$

$$\mathbf{d}_{\text{ACC}}(X' \to X) = \frac{\mathbf{d}_{\text{ACC}}(X') - \mathbf{d}_{\text{ACC}}(X)}{\mathbf{d}_{\text{ACC}}(X)}$$

AGREEMENT This metric measures the extent to which the model's predictions vary in response to different interventions. The rationale behind this metric stems from the recognition that the task under consideration entails a multiple-choice problem. Additionally, in real-world scenarios, it is often the case that a definitive 'ground truth' may not exist. Consequently, we employ this metric to measure the divergence of predictions. This metric goes beyond simple accuracy, which merely determines the correctness or incorrectness of predictions. Instead, this objective is to evaluate the diversity of predictions, thereby taking into consideration the range of

errors that may arise. To calculate the **AGR** score, which is a modification of Fleiss' kappa (Fleiss and Cohen, 1973), we begin with a list of N names and obtain a score:

$$\mathbf{d}_{\text{AGR}}(X) = \frac{1}{|N| \cdot |N - 1|} \sum_{j=1}^{k} (n_j \cdot (n_j - 1))$$

$$\mathbf{d}_{\text{AGR}}(X' \to X) = \frac{\mathbf{d}_{\text{AGR}}(X') - \mathbf{d}_{\text{AGR}}(X)}{\mathbf{d}_{\text{AGR}}(X)}$$

where $|N|$ indicates the total number of names in name lists, k the number of categories (e.g. in our case, $k = 3$, {(a),(b),(c)}), and n_j the number of instances predicting the answer as category j. The **AGR** score ranges from 0 to 1, with a score of 1 indicating complete agreement among all name instances in their category prediction, and a score of 0 indicating no agreement. This metric enables us to assess the degree to which a model's predictions are sensitive to different interventions.

4 Explanations of Causal Effects

The causal analysis shows the surface-level comparison of model outputs but fails to capture the nuanced processes underlying each model's reasoning. By probing the internal workings of the models, we seek to gain insights into how the models derive their conclusions and also where their approaches diverge. We use two approaches to gain a deeper understanding of the models' predictions. First, we analyze the models' internal representations to discern how they encode various names. Specifically, we focus on the distinction in contextualization between the embeddings of frequent names and less frequent names. Second, we apply a diagnostic technique based on neuron activation to pinpoint how the models process names.

4.1 Contextualization of Name Representations

We investigate the contextualization of name representations in language models with respect to their characteristics. We partition the names based on frequency MOST and LEAST and compare the degree of contextualization. To be specific, we measure the similarity between name representations at each layer of the model by following the approach proposed by Wolfe and Caliskan (2021). In order to ensure that the embeddings being compared are based on the same space, we restrict the comparison to representations within each layer and do not compare across different layers. We adopt two commonly used metrics to validate the overall trend observed in our analysis.

COSINE SIMILARITY The cosine-similarity of name w, in layer l is formalized as followes:

$$c(\mathbf{w})_l = \frac{1}{n^2 - n} \sum_i \sum_{j \neq i} cos(\vec{w}_i, \vec{w}_j)$$

where n refers to the total number of name pairs. This corresponds to the self-similarity studied in (**?**Wolfe and Caliskan, 2021). The measure lies ranges from 0 to 1, where 1 indicates high similarity, and 0 otherwise.

LINEAR CKA (Centered Kernel Alignment) This similarity metric measures similarity in neural network representations and was proposed by Kornblith et al. (2019). It ranges from 0 to 1, where 1 indicates perfect similarity, and 0 otherwise.

$$\frac{||\mathbf{x_j}^\top \mathbf{x_i}||_F^2}{||\mathbf{x_i}^\top \mathbf{x_i}||_F ||\mathbf{x_j}^\top \mathbf{x_j}||_F}$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ indicates two randomly selected name embeddings, such that $i \neq j$.

4.2 Neuron Activations

Previous work has explored the activation patterns of neurons in deep neural networks for the domains of language and vision as a means of gaining insight into the inner workings of such networks (Karpathy et al., 2015; Poerner et al., 2018; Olah et al., 2018; Dalvi et al., 2019). It has been demonstrated that the feed-forward network (FF) component of transformer architectures encodes a significant amount of information (Wang et al., 2022; Geva et al., 2021). Building on this prior work, we conducted a detailed analysis of how neuron activations vary according to different characteristics of the input data. Our analysis involved extracting the activations of the FF network's neurons based on the hidden states of previous layers and applying non-negative matrix factorization (NMF) (Cichocki and Phan, 2009) to decompose these activations into semantically meaningful components. By visualizing groups of neuron activations, we aim to gain a better understanding of the models' internal mechanisms, and how the models construct their representations and predictions. For the detailed algorithm see Appendix B outlines the steps involved in this analysis.

Effect size: $\mathbf{d}_{ACC}$ (Accuracy)	Not-finetuned (Epoch 0)			Fine-tuned (Epoch10)		
	GPT2	BERT	ROBERTA	GPT2	BERT	ROBERTA
MOST → LEAST	-.07 (.354)	**.258***** (<.001)	-.04 (.534)	.002 (.956)	.007 (.841)	.004 (.884)
MALE → FEMALE	.001 (.801)	.005 (.634)	-.025 (.627)	.002 (.819)	-.002 (.965)	.002 (.751)
MOST MALE → LEAST MALE	-.059 (.365)	**.275***** (<.001)	-.018 (.627)	-.004 (.906)	.006 (.885)	.011 (.751)
MOST FEMALE → LEAST FEMALE	-.089 (.349)	**.241***** (<.001)	-.06 (.800)	.008 (.990)	.008 (.800)	-.002 (.954)

Table 1: Direct Effect: Accuracy ($\mathbf{d}_{ACC}$) score of the models with and without fine-tuning. The numbers in parentheses are *p-values*. The values in bold indicate the significant effects with *p-values*< 0.05. The results show that after fine-tuning, the direct effects are not significant.

Effect size: $\mathbf{d}_{AGR}$ (Agreement)	Not-finetuned (Epoch 0)			Fine-tuned (Epoch10)		
	GPT2	BERT	ROBERTA	GPT2	BERT	ROBERTA
MOST → LEAST	-.0004 (.954)	**.058***** (<.001)	**.048***** (<.001)	**.013*** (.02)	**.022***** (<.001)	**.012*** (.02)
MALE → FEMALE	.02 (.712)	.009 (.306)	.010 (.267)	.004 (.565)	-.002 (.722)	.007 (.354)
MOST MALE → LEAST MALE	.003 (.748)	**.068***** (.0)	**.060***** (<.001)	**.017*** (.028)	**.027***** (<.001)	.015 (.052)
MOST FEMALE → LEAST FEMALE	-.004 (.691)	**.047**** (.004)	**.03***** (<.001)	.009 (.262)	.016 (.240)	**.010*** (.036)

Table 2: Direct Effect: Agreement ($\mathbf{d}_{AGR}$) score of the models with and without fine-tuning. The numbers in parentheses are *p-values*. The values in bold indicate the significant effects with *p-values*< 0.05. The results show that after being fine-tuned, the effects show significance in the frequency of the names (row1). The asterisks indicate the significance level: ($^{***}p \leq 0.001, ^{**}p \leq 0.01, ^{*}p \leq 0.05$)

5 Experimental Setup

Dataset We use the SOCIALIQA dataset from Sap et al. (2019). The selection of this dataset is motivated by its suitability for investigating model behavior in a social context, as the dataset consists of questions for probing *emotional* and *social* intelligence in everyday situations. By analyzing the model's responses to questions pertaining to social and emotional intelligence, valuable insights can be gleaned regarding the models' handling of some nuances of human behavior. Since the dataset is based on a social setting, it would be misleading if the models yielded different predictions based on different names. To construct the template **T**, we used the AllenNLP coreference resolution model (Gardner et al., 2018), which has high performance[3]. This model is used to detect named entities and resolve their corresponding pronouns, facilitating the construction of templates for our experiments.

Names List We use U.S. census names dataset[4], following (Mehrabi et al., 2020) to intervene the name placeholders. It contains 139 years of U.S. census baby names, their corresponding gender, and respective frequencies. To form intervention name lists based on frequency, we filtered out the most frequent k names over all years for N_{MOST}, and the least frequent k names over all years for N_{LEAST}. We set $k = 200$.

Model We use three widely used models, GPT2 (Radford et al., 2019), BERT (Devlin et al., 2019), and ROBERTA (Liu et al., 2019). We customized each model with a linear layer

---

[3]F1 score 80.2 on CoNLL benchmark dataset

[4]http://www.ssa.gov/oact/babynames/names.zip

		Not-finetuned (Epoch 0)			Finetuned (Epoch10)		
		GPT2	BERT	ROBERTA	GPT2	BERT	ROBERTA
Indirect Effect	MOST	0.055	0.107	0.074	0.052	0.047	0.037
	LEAST	0.043	0.091	0.171	0.053	0.039	0.031
	FEMALE	0.072	0.145	0.185	0.079	0.063	0.051
	MALE	0.030	0.059	0.034	0.0260	0.025	0.018

Table 3: Indirect Effect of name lists across models. The results show that relative to Non-finetuned models, the indirect effect of names on predictions is marginally reduced in fine-tuned models.

on top to perform a multiple-choice selection task. The feed-forward (FF) linear layer was obtained by logits $=$ **Model**(X), $\hat{y} =$ **FF**(logits). The hyper-parameter setting for the training is described in Appendix A.

6 Results and Discussion

6.1 Direct Effect

ACCURACY The results of the direct effect of accuracy for different sets of interventions are presented in Table 1. Comparing the first three columns (*not-finetuned*) with the subsequent three columns (*fine-tuned)*, we observe that the causal effect of accuracy is not statistically significant when the models are fine-tuned. This trend holds consistently true across all three models examined in this study. This suggests that the direct effect of name characteristics on accuracy is not significant when fine-tuned. The effect sizes of the *not-finetuned* models are reported in accordance with previous literature that predominantly focuses on these models (Wolfe and Caliskan, 2021; Shwartz et al., 2020). However, it is crucial to emphasize the efficacy of fine-tuning, as it reflects a more realistic scenario for model deployment (Jeoung and Diesner, 2022). We compared the effect sizes of the not-finetuned models with those of the fine-tuned models, thereby examining the impact of fine-tuning on model behavior. We also provide an analysis of the correlation between the model's accuracy and effect sizes in Appendix D.

AGREEMENT The analysis of the direct causal effect of agreement (d_{AGR}) shows that a significant difference in name lists based on frequency persists even after fine-tuning all three models (Table 2, first row). This suggests that despite the fine-tuning process, the models continue to exhibit variations in their agreement on predictions based on the frequency of names used.
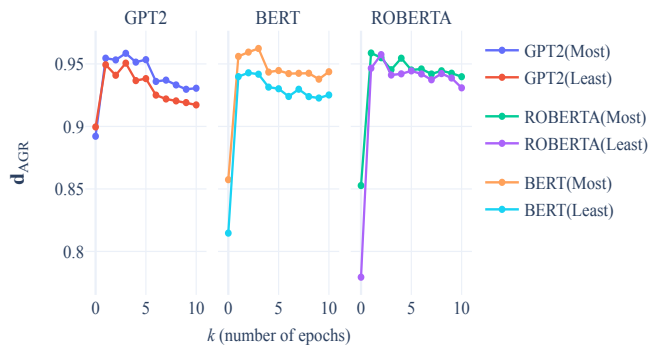

Figure 2: The d_{AGR} of MOST and LEAST values over the training phase (number of epochs). For GPT2 and BERT, the gap of MOST values and LEAST is consistent across the number of epochs.

Specifically, the positive and significant value of MOST $\rightarrow$ LEAST indicates that the prediction is more divergent forLEAST than MOST. This implies that when the model makes incorrect predictions, the resulting predictions tend to be more inconsistent or diverse, rather than consistent.

Figure 2 illustrates the disentangled values for d_{AGR} across different epochs during the training phase. For both GPT2 and BERT, a consistent gap between MOST and LEAST is observed throughout the training epochs. In contrast, for ROBERTA, although the gap is not consistent across all epochs, the agreement measures for MOST remain consistently higher than those for LEAST. This discrepancy in the gap between ROBERTA and the other models could potentially be attributed to the robust optimization design of ROBERTA, which complements that of BERT (Liu et al., 2019). Also, these findings are consistent with the conclusion drawn by (Basu et al., 2021), who also observed that ROBERTA generates the most

robust results. Overall, the findings indicate that the agreement ratio of LEAST consistently remains lower than that of MOST throughout the training phase, suggesting that the predictions for LEAST are more divergent.

6.2 Indirect Effect

Table 3 presents the results pertaining to the indirect effect of name lists on predictions. Specifically, the indirect effect quantifies the sensitivity of pronouns associated with names on model predictions. Overall, the findings indicate that, in comparison to non-finetuned models, the indirect effect of names on predictions is marginally reduced in fine-tuned models. For BERT and ROBERTA, the indirect effect of both frequency and gender is diminished when finetuned. However, for GPT2, the indirect effect is reduced in most cases, except for the name lists of LEAST and FEMALES.

6.3 Contextualization Measures

In order to gain insight into how names are internally contextualized in the transformer models, we conducted a preliminary analysis of name representations. To do so, we extracted the embeddings of N_{MOST} and N_{LEAST} samples from fine-tuned GPT2 and measured their similarity. The results are presented in Figure 3 and 4. The SELF-SIMILAR(Most) and SELF-SIMILAR(Least) measures represent the similarity between the MOST and LEAST names, respectively, while the INTER-SIMILARITY(Most-Least) measure quantifies the similarity between the Most and Least names. The trends observed for both CKA and cosine similarity measures are similar, although with different magnitudes (details of these metrics are discussed in section 4). These consistent trends are robust across different evaluation metrics. The results show that in the first two layers, the similarity scores are low, but they increase across the mid-layers. However, in the last layer, the similarity of the embeddings of LEAST names is lower compared to MOST names. This finding partly explains Table 2 first row, which indicates the fine-tuned GPT2 has a significant direct effect on the agreement measure on MOST and LEAST. The relatively low similarity of the embeddings of LEAST names shows that they exhibit higher variability, being less contextualized compared to that of MOST.
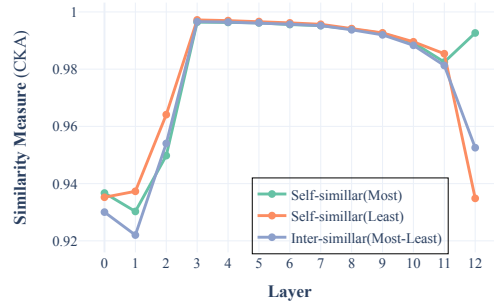

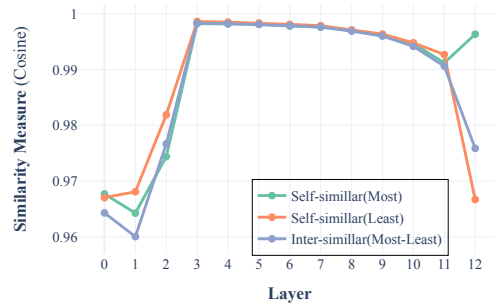Figure 3: CKA measures across layers

Figure 4: Cosine similarity measures across layers

6.4 Neuron Activations

To further investigate the differences in neuron activations, we conducted an analysis using GPT2 fine-tuned model. The results of this analysis are presented in Table 4, where each color represents the components of the neurons that got activated. These components correspond to the clusters obtained from the non-negative factorization on feedforward neurons. Our observations indicate that less frequent names exhibit two distinct behaviors: 1) they are sub-tokenized into two or more tokens, and 2) they are not activated by the same neuron components as the frequent names. This analysis does not provide an explanation for the *cause* or *reason* for the divergent predictions but rather sheds light on the internal behavior of the model, namely how the neurons activate, which may be related to the divergent predictions observed for the least frequent names.

6.5 Mitigating Strategy: Data Augmentation

Our findings suggest that incorporating a more diverse set of first names into the training data can serve as a potential approach to mitigate the di-

Examples
Mary was always the type who liked to party , she was excited it was her birthday , and invited people to her house . [SEP] [SEP] why did Mary do this ? loved the party scene [PAD] [PAD] social ize [PAD] [PAD] was not the party girl [PAD] [PAD] [PAD]
Elizabeth was always the type who liked to party , she was excited it was her birthday , and invited people to her house . [SEP] [SEP] why did Elizabeth do this ? loved the party scene [PAD] [PAD] social ize [PAD] [PAD] was not the party girl [PAD] [PAD] [PAD]
James was always the type who liked to party , he was excited it was his birthday , and invited people to his house . [SEP] [SEP] why did James do this ? loved the party scene [PAD] [PAD] social ize [PAD] [PAD] was not the party boy [PAD] [PAD] [PAD]
Robert was always the type who liked to party , he was excited it was his birthday , and invited people to his house . [SEP] [SEP] why did Robert do this ? loved the party scene [PAD] [PAD] social ize [PAD] [PAD] was not the party boy [PAD] [PAD] [PAD]
And rine was always the type who liked to party , she was excited it was her birthday , and invited people to her house . [SEP] [SEP] why did And rine do this ? loved the party scene [PAD] [PAD] social ize [PAD] [PAD] was not the party girl [PAD] [PAD] [PAD]
Le u ven ia was always the type who liked to party , she was excited it was her birthday , and invited people to her house . [SEP] [SEP] why did Le u ven ia do this ? loved the party scene [PAD] social ize [PAD] [PAD] was not the party girl
Nav ajo was always the type who liked to party , he was excited it was his birthday , and invited people to his house . [SEP] [SEP] why did Navajo do this ? loved the party scene [PAD] [PAD] social ize [PAD] [PAD] [PAD] was not the party girl [PAD] [PAD] [PAD]
Wind field was always the type who liked to party , he was excited it was his birthday , and invited people to his house . [SEP] [SEP] why did Wind field do this ? loved the party scene [PAD] [PAD] social ize [PAD] [PAD] was not the party girl [PAD] [PAD] [PAD]

Table 4: Neuron Activation analysis. The section above lists the examples of Frequent Names: *Mary, Elizabeth, James, Robert* while the section below shows the examples of Least Frequent Names: *Andrine, Leuven, Navajo, Windfield*. The color corresponds to the group of components of the neurons that are activated.

vergent behavior of language models. Among all first names in the SOCIALIQA training dataset, we observed around 66% of first name instances represent the 10% of the most frequent first names in the U.S. Census data. In terms of frequency, these names account for 97% of all first-name instances in the training dataset (Fig in Appendix C). Such skewed yet highly likely distributions of demographic information in the training dataset may inadvertently introduce biases in the model outputs, as evidenced by previous studies (Buolamwini and Gebru, 2018; Karkkainen and Joo, 2021). To address this issue, recent research by (Qian et al., 2022) has demonstrated that augmenting the training data with diverse social demographics can lead to improved model performance and robustness.

7 Related Work

Previous research has shown that pre-trained language models are susceptible to biases related to people's first names, e.g., in the contexts of sentiment analysis (Prabhakaran et al., 2019) and text generation (Shwartz et al., 2020). Wolfe and Caliskan (2021) demonstrated that less common names are more likely to be subtokenized and associated with negative sentiments compared to frequent names. In our work, we further extended this prior work by analyzing the impact of fine-tuning models on first names adopting the causal framework.

A growing body of research has explored the incorporation of causality in language models. For instance, Feder et al. (2021) proposed a causal framework by incorporating additional fine-tuning on adversarial tasks. Similarly, Vig et al. (2020) demonstrated the use of causal mediation on language models to mitigate gender bias. Unlike Vig et al. (2020), our approach focuses on applying causal analysis in the input sequence space and exploring the causal relationships between input sequence components and model predictions.

8 Conclusion

In this paper, we introduced a controlled experimental framework to assess the causal effect of first names on commonsense reasoning. Our find-

ings show that the frequency of first names exerts a direct impact on model predictions, with less frequent names leading to divergent outcomes. We suggest careful consideration of the demographics in dataset design.

9 Broader Impact

The data used in our analysis contains no private user information. As for ethical impact, the systematic experimental design we used provides an approach for conducting controlled experiments in the context of natural language processing research, particularly with a focus on the influence of first names on language models.

10 Limitation

Our investigation focuses on one aspect of commonsense reasoning restricted to one dataset. There may be numerous other factors in real-world applications. Therefore, our findings may not comprehensively capture the entirety of commonsense reasoning phenomena. Another limitation is that for the sake of simplicity and feasibility, we assumed a fixed threshold of k=200 to categorize frequent and less frequent names. However, this threshold may not be universally applicable to all contexts or datasets, and different thresholds could lead to different results.

References

Jing Bai, Jian-Yun Nie, and François Paradis. 2004. Using language models for text classification. In *Proceedings of the Asia Information Retrieval Symposium, Beijing, China*. Citeseer.

Priyam Basu, Tiasa Singha Roy, and Ashima Singhal. 2021. But how robust is roberta actually?: A benchmark of sota transformer networks for sexual harassment detection on twitter. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 1328–1333. IEEE.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Andrzej Cichocki and Anh-Huy Phan. 2009. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721.

Fahim Dalvi, Avery Nortonsmith, Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019. Neurox: A toolkit for analyzing individual neurons in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9851–9852.

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.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

Sullam Jeoung and Jana Diesner. 2022. What changed? investigating debiasing methods using causal mediation analysis. *arXiv preprint arXiv:2206.00701*.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

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. *arXiv preprint arXiv:1907.11692*.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 231–232.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill*, 3(3):e10.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 373–392.

Thorsten Brants Ashok C Popat Peng and Xu Franz J Och Jeffrey Dean. 2007. Large language models in machine translation. *EMNLP-CoNLL 2007*, page 858.

Nina Poerner, Benjamin Roth, and Hinrich Schütze. 2018. Interpretable textual neuron representations for nlp. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 325–327.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. 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 5740–5745.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. 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 4463–4473.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. " you are grounded!": Latent name artifacts in pre-trained language models. *arXiv preprint arXiv:2004.03012*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Fei-Yue Wang, Kathleen M Carley, Daniel Zeng, and Wenji Mao. 2007. Social computing: From social informatics to social intelligence. *IEEE Intelligent systems*, 22(2):79–83.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532.

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

A   Training Hyperparameters

For the train/test split, we followed the original split provided by the data source (Sap et al., 2019). The hyper-parameters used for training are as follows: *AdamW* optimizer, with learning rate *1e-5*, 10 epochs. The checkpoints were saved at the end of every epoch.

B   Neuron Activation Analysis

Algorithm 1: Neuron Activation Analysis

Data: $\mathbf{X} := (x_1, x_2, \ldots, x_n)$, n tokens
Result: $\mathbf{M} \in \mathbb{R}^{k \times n}$, k components
$L \leftarrow$ # layers;
for $i \leftarrow 1$ **to** L **do**
  $\mathbf{X}' \leftarrow f\mathbf{block}_i(\text{pre-mlp}_i(\mathbf{X}))$;
  $\mathbf{y}_i \leftarrow f\mathbf{block}_i\text{mlp}(\mathbf{X}')$
end
$\mathbf{Y} \leftarrow concat(\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_L) \in \mathbb{R}^{L \times h \times n}$;
$\mathbf{M} \leftarrow \text{NMF}(\mathbf{Y})$

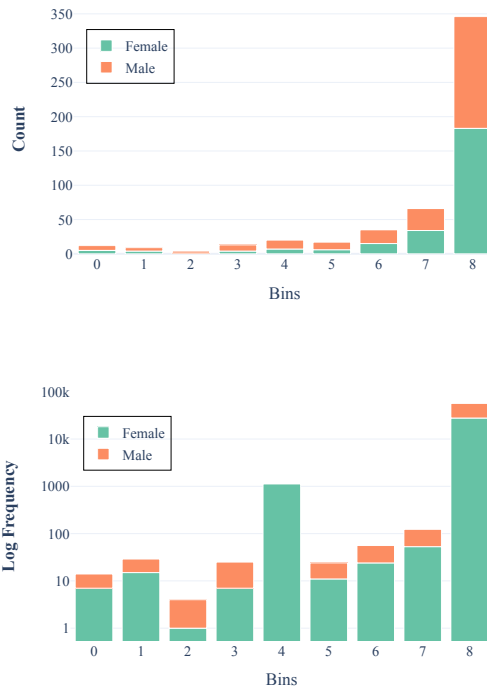---

C   SOCIALIQA train set names configuration

Figure 5: Distribution of first names in the train split in SOCIALIQA dataset. The first names are sorted in ascending order based on U.S. census data frequency and filled into the bins based on quantiles. The x-axis represents the Bins. (Above) displays the count of the first names that fall into those bins, showing the prevalence of first names based on whether they are used in the training set of not (Below) shows the frequency of these names in the dataset on a logarithmic scale along the y-axis, showing how frequently these names appear in the dataset.

Effect size:	$\mathbf{d}_{ACC}$			$\mathbf{d}_{AGR}$		
Model	GPT2	BERT	ROBERTA	GPT2	BERT	ROBERTA
MOST → LEAST	.473 (.142)	-.427 (0.19)	-.109 (.75)	.536 (.089)	-.500 (.117)	.500 (.17)
MALE → FEMALE	.045 (.894)	-.264 (.433)	.055 (.873)	-.555 (.077)	-.55 (.077)	-.591 (.056)
MOST MALE → LEAST MALE	.264 (.433)	-**.609*** (.047)	-.073 (.832)	.473 (.142)	-**.645*** (.032)	-.6 (.051)
MOST FEMALE → LEAST FEMALE	**.618*** (.043)	-.264 (.433)	.191 (.574)	.391 (.235)	-.418 (.201)	-.145 (.67)

Table 5: Spearman Correlation between Model's Accuracy and Effect Size: The values show the Spearman's Correlation between the model's accuracy with the effect size ($\mathbf{d}_{Acc}$ and $\mathbf{d}_{AGR}$). The numbers in parentheses indicate the *p-values*. The values in bold indicate the statistical significance with *p-values* < 0.05. The results show that in most cases, the correlation values are not statistically significant.

D Accuracy and Effect Size Correlation analysis

The relationship between the effect size and the model's performance, measured by accuracy, was investigated in order to determine whether there was any correlation. Table 5 presents the correlation analysis between the model's accuracy and two corresponding effect sizes, namely ($\mathbf{d}_{ACC}$, and $\mathbf{d}_{AGR}$). Specifically, for each epoch during the fine-tuning phase, the model's accuracy and effect sizes were compared, and Spearman's correlation coefficient was computed. The results indicate that, in most cases, the correlation values were not statistically significant (*p values* ≤ 0.05). This suggests that there is no significant association between the improvement in model accuracy and corresponding effect sizes, either positive or negative. By examining the raw data, it was observed that while the models' accuracy increased, the effect sizes remained relatively constant (as shown in Fig 2) throughout some points of the epoch, indicating that there exists some bottleneck in fine-tuning process, as the effect sizes were not effectively mitigated even with the improvement in accuracy.