Involving Affected Communities and Their Knowledge for Bias Evaluation in Large Language Models

Vildan Salikutluk
vildan.salikutluk@tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Elifnur Doğan
elifnur.dogan@stud.tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Isabelle Clev
isabelle.clev@stud.tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Frank Jäkel
frank.jaekel@tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

ABSTRACT

Large Language Models (LLMs) often show various types of biases, e.g., with regard to gender or religion, which was already documented in previous research. However, to further examine such biases in a human-centered fashion, we conduct a survey of the affected community, i.e. in this case the Muslim community, to learn about their perspectives, expectations, and opinions on LLM-based systems and their possible applications. We find that the participants assume that their name is one of the most important factors based on which LLMs might assess them unfairly. This concern is confirmed by the results of an evaluation in which we test several state-of-the-art LLMs (GPT-3.5, GPT-4, Llama 2 and Mistral AI) as all of them display biases against Muslim names. We find intersectional biases, as female and male Muslim names are assigned in differently skewed ways by the LLMs. We also collected common and uncommon names from the Muslim community, which allowed us to test whether the likely representation of the names in the training data influences how the outputs are produced. In fact, the results demonstrate differences in LLM outputs for common and uncommon names. We show that involving affected communities and their intuitions and knowledge allowed us to investigate a factor, i.e. names, that is not only important to them but also can be used to uncover biases in LLMs in meaningful ways.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → **Empirical studies in HCI**; **Interactive systems and tools**.

KEYWORDS

Large Language Models, Biases, Muslim, Names

1 INTRODUCTION

Systems based on artificial intelligence (AI) and specifically Large Language Models (LLMs) can potentially be helpful for solving important problems in many areas, e.g. in medicine [7, 36], education [7, 27], business [19], or to combat climate change [11, 34]. However, applying LLMs can also cause harm and exacerbate (existing) inequities [7, 8, 43]. While LLMs and systems like, e.g., ChatGPT, demonstrate a wide range of capabilities that can be useful, it is also apparent that aside from their potential for unintended misuse [19] there are a number of issues within the models themselves. In particular, LLMs exhibited biases about, e.g., race [12, 18], religion [1, 9, 24], gender [30], disability [26], and more. It appears that OpenAI for instance manages to mostly prevent ChatGPT from producing explicitly violent and toxic outputs – which was the case in previous versions [1] – with the help of reinforcement learning with human feedback [2, 33]. Thus, while LLMs often refuse to provide an answer for questions that explicitly include protected features, e.g., race, religion, etc., it was also repeatedly shown that such filters can be bypassed in several ways [16, 42]. Moreover, recent work has also shown that implicit biases and harmful stereotypes are still present in most state-of-the-art LLMs [6]. Using indirect prompting methods or proxy variables, such as names, can elicit underlying biases that are still present in many LLM-based systems [6, 41]. Furthermore, names can be used to investigate intersectional biases [14], i.e. when multiple biases occur at once because a person is part of more than one marginalized group, which can exacerbate negative outcomes [32, 35]. The associations, stereotypes and possible biases that such systems show for names are particularly interesting to investigate further as names are relevant in many down-stream applications in which LLMs might be used. In general, if decision support systems are used in high stakes decision-making contexts, their intrinsic biases can have serious negative impacts on people's lives [4, 37, 43]. Names could be seemingly irrelevant data that still trigger biased responses in LLMs as they can be (implicit) signifiers of protected features such as a person's nationality, race, gender or religion, and also represent intersections between them. Thus, we investigate whether the underlying associations of recent state-of-the-art LLMs are skewed and display a tendency for negative stereotyping based on names. Specifically, we look at the intersectional factors of gender and religion. We examine whether there is a difference in how often LLMs assign certain positive and negative roles to male and female Muslim vs. non-Muslim names. In addition, we also prompt the LLMs with both common and uncommon names because it is likely that the frequency of the name has an effect on both the original training of the model and any subsequent debiasing efforts. We conduct a survey of Muslims (in mostly Western countries, specifically Germany) to collect both common and uncommon Muslim names and also ask participants

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

about their attitudes, expectations and opinions of LLMs. With the collected names, we evaluate state-of-the-art LLMs in several settings to test whether they produce biased outputs based on different types of names.

2 RELATED WORK

Religion has already been investigated in previous work, and LLMs were often shown to be perpetuating harmful stereotypes against Muslims and Islam [1, 12, 24, 25, 35]. There is also previous work on gender biases in LLMs showing, e.g., how female and male characters are described in terms of stereotypes by a variety of LLMs [30, 41]. Furthermore, several works [29, 31, 35, 39] show that currently established methods for debiasing LLMs are considerably less effective when it comes to intersectional biases, and even models which displayed decent fairness levels in regard to individual demographics were much less fair for the intersections of them. While these studies are more inclusive by covering multiple biases and their overlaps, many of them still lack nuances and many intersectional dimensions are yet to be explored [22, 23, 35]. While many studies show that explicit biases (intersectional or not) are present in LLMs, i.e. skewed and harmful output due to naming protected features, there is also recent work investigating associations and choices of LLMs between two options based on implicit biases [6]. In other work, word embeddings have been used to measure implicit biases [13], but these embeddings are not accessible in many of the state-of-the-art LLMs such as, e.g., ChatGPT. Therefore, Bai et al. [6] show that several state-of-the-art LLMs display biases in their choices even if they are explicitly debiased by utilizing a modified version of the implicit associations test [21] showing, e.g., that GPT-4 disagrees with blatant statements such as “women are bad at managing people” but has no problem with readily choosing Ben over Julia if asked which one of them should lead a management workshop. These types of choices were posed to several LLMs in their study, and their results show that there are strongly implicit negative biases and thus skews in LLM systems’ choices in terms of gender, race, and other protected features.

3 SURVEY ABOUT EXPECTATIONS AND KNOWLEDGE ABOUT LLMS IN THE MUSLIM COMMUNITY

In a survey, we first obtain Muslim names and judgements about their frequency. Second, we collect data about the Muslim participant’s attitude towards LLM-based AI systems.

3.1 Methods

We conducted an online survey in which 77 Muslims participated. Available languages were English, German and Turkish (3 participants used the English version of the survey, 87 the German, and 7 the Turkish version). We distributed the survey mostly through (social media) platforms in Germany. In terms of demographics, we solely asked participants whether they identify as Muslim or not. The experiment was approved by the local ethics board and all participants provided informed consent. In the first part of the online survey, participants were asked to provide common and uncommon Muslim names based on their subjective assessment(s), subdivided into female and male. Subsequently, participants were then asked to evaluate the perceived frequency of 23 (pre-determined) Muslim names, by indicating whether the name is "common" or "uncommon" (there was also a third "I don’t know" option). In the second part of the survey, participants were asked to first indicate their familiarity with LLM-based AI systems, such as ChatGPT, with questions about whether they know such systems and whether they used them before. Then participants were presented with a short description of a scenario and were instructed to answer questions about it. The presented scenario and instructions were the following: "This study is about your attitude towards the use of artificial intelligence (AI). For example, AI systems might automate some tasks (in the future). Suppose an AI system (similar to, e.g., ChatGPT) reads your application for a job and evaluates it to decide whether you get the position. Please indicate your answer to the following statements that relate to the scenario described above. Please remember that there is no right or wrong answer, it is all about your personal opinion. If you want to, you can give reasons for your assessment after each question." The participants were then asked to rate on a 3-point scale whether they think that such a system would judge a job application of theirs more fairly (=1), the same (=2), or more unfairly (=3) compared to a human. Furthermore, participants were asked to indicate whether they think that certain aspects would influence the evaluation of their job application by an AI system. Specifically, we asked the participants to rate whether their name, age, place of birth, work experience, education, spoken languages, certifications or skills would influence the AI system’s evaluation positively (=3), not at all (=2) or negatively (=1). Additionally, participants answered 12 (adjusted) items from the Trust in Automation questionnaire [28]. We instructed them to answer the question also in reference to the described scenario and system above. For all questions in the second part of the survey, participants could state their reasons for an answer as free form comments.

3.2 Results

Overall, in the first part of the survey, participants listed 560 common female Muslim names, 669 common male Muslim names, 277 uncommon female Muslim names and 270 uncommon male Muslim names. We additionally asked our participants to rate a list of common (12 female and male) and uncommon (9 uncommon female and male) names we provided. All ratings were in line with what we excepted each name to be rated as. In Table 1, there are examples of names we asked participants to rate in terms of their frequency and in the "Rating" columns are the percentages of how many of the participants agreed on the name being either common or uncommon, respectively.

In the second part of the survey, all participants indicated both knowing LLM-based systems, like ChatGPT, and also having used them. When we asked them to rate how an LLM-based AI system, such as the one we described, would assess their job application, 22.0% indicated that they think the system would be fairer than a human, 19.5% that the AI system would be more unfair than a human and 5.2% that it would be the same. The rest of the participants indicated that they don’t know. We further analyzed all written responses (29) that the participants gave as to why they choose their answers. 11 participants who would expect the AI to

		Uncommon				Common			
M-F	**Name**	Atikah	Esila	Efnan	Hifa	Zeynep	Aisha	Fatima	Maryam
	SR	8.390.000	2.680.000	315.000	2.440.000	109.000.000	108.000.000	316.000.000	145.000.000
	Rating	92.2%	72.7%	90.9%	85.7%	97.4%	97.4%	98.7%	97.4%
M-M	**Name**	Ecir	Hasbi	Benan	Bukra	Mohammed	Ahmed	Omar	Hassan
	SR	47.900.000	7.390.000	34.900.000	7.960.000	6.920.000.000	3.140.000.000	7.080.000.000	1.320.000.000
	Rating	87.1%	84.4%	94.8%	72.7%	97.4%	98.7%	96.1%	96.1%
NM-F	**Name**	Avalee	Aurela	Aviana	Elja	Amy	Lisa	Emily	Elizabeth
	SR	1.460.000	1.750.000	583.000	1.740.000	1.460.000.000	2.300.000.000	2.220.000.000	1.810.000.000
NM-M	**Name**	Arlo	Lenn	Vinny	Yorick	Michael	Peter	John	Justin
	SR	47.900.000	7.730.000	34.900.000	7.960.000	6.920.000.000	3.140.000.000	2.300.000.000	1.320.000.000

Table 1: Overview of used Muslim and non-Muslim names in the evaluation of LLMs. M-F = Female Muslim names, M-M = Male Muslim names, NM-F = Female non-Muslim names, NM-M = Male non-Muslim names. The number of Google search results for each name are presented in rows "SR". The percentage of how uncommon or common each name was perceived to be in our survey is presented as "rating", e.g., the name "Atikah" was rated as uncommon by 92.2% of our Muslim participants.

be fairer commented they hope the AI would only judge their qualifications and not their names, e.g. one participant writes "[There's] no discrimination based on names, [and] more focus on quality and experience."[1] Three participants who indicated the AI to be no different from humans in assessing their applications all indicated that this is due to the training data, e.g. "AI systems get their information from humans, so it's the same."[1] This was also the overall sentiment of the 10 participants, who indicated that they do not know whether an AI system would be more or less fair compared to humans in assessing their application. Five participants indicated that the AI would be less fair towards their application than a human, as the AI cannot judge, e.g., social features such as "[...] if someone fits into the team as a person."[1] Further, we analyzed how participants rated different aspects, i.e., their name, age, place of birth, work experience, education, languages, certifications, and skills to be influencing the assessment of their application if an AI system (such as the one described) would evaluate it. From all aspects we queried, we find that the participants expect their names to be influencing the AI systems most negatively, with mean value 1.69 (SD = 0.73). We use pairwise t-test comparisons with Bonferroni adjusted alpha levels of .00714 per test (.05/7) and find that their name is rated as influencing the evaluation of their application in a significantly more negative fashion compared to all other aspects, except for age and birthplace (all p-values < 0.001). This sentiment is also reflected in the participants' comments, e.g. "No question, my name is not going to pass through the AI's filters."[1] Finally, the answer to our adjusted version of the Trust in Automation questionnaire [28] showed most agreement given to items which refer to the system making errors, and being generally cautious towards unfamiliar automated systems. The item with the lowest mean value of agreement was "The developers take my well-being seriously." Furthermore, participants on average rather disagreed when they are asked whether they would trust and rely on such a system.

[1]translated from German

4 EVALUATION OF POTENTIAL NAME-BASED BIASES IN LLMS

While explicit mentioning of protected variables often leads to filtered responses in LLMs, they still exhibit implicit stereotypical and harmful associations [6]. Thus, we use the names from Table 1 to test potential biases of the LLMs' when they assign Muslim and non-Muslim names to roles with positive and negative connotations.

4.1 Methods

To evaluate LLMs, we present a story and let each LLM assign names to each of the roles in the story from a list we provide to test whether assignments of names are biased. We test a variety of state-of-the-art LLMs, specifically the following models: OpenAI's GPT-3.5 and GPT-4, Meta's Llama-2-7b-chat, and Mistral AI's Mistral-7B-Instruct-v0.2. We test these models in different settings, such as a police setting in which there are two police officers and one suspect, with the role of "suspect" having negative connotations. It is even specified in the story that the suspect is in fact guilty. Next, we test a court setting, in which there are two prosecutors and one defendant, with the role of "defendant" having negative connotations as it is specified that the defendant is guilty of the accused crime. We also test a job setting which includes one interviewer and two applicants for the same position. Here, we do not check for the assignment to a role with a negative connotation, but rather which of the applicants receives the job offer. Lastly, there is a "neutral" retail setting where there are two roles of retail workers and one customer as "control" setting with no positive or negative connotations for any role. The LLMs are tasked to "fill in the blanks". They are given settings with set characters in certain roles and have to assign names to them. Particularly, we provided name lists to the LLMs and let them generate variable assignments for each of the marked roles from the list. The roles in the given setting always have variables X, Y and Z with X and Y being the roles that were assigned twice and Z being the role that was assigned once. We test every possible combination of features pairwise and thus vary names with respect to all features, i.e. whether a name is Muslim or not as well as whether it is common or uncommon and test all combinations for female and male names separately. For each of the comparisons,

we provided a name list containing eight names overall, with four names from each of the compared feature combinations. Overall, we let each LLM produce answers 50 times to the prompts for each of the settings and each combination of features. As there are many factors and uncertainties to consider when judging whether a name is common, we use two measures. First, in our survey, Muslim participants indicated whether pre-determined Muslim names are common or not. Additionally, for both the Muslim and non-Muslim names, we used the number of search results on Google as a proxy for how common they are and how frequent the names might be in the LLMs' training data. If the number of results for the search of a specific name is very high (compared to the other names), we assume it to be common. If the number is low, we assume it to be uncommon. An overview of all names we used in our evaluation with their number of search results are shown in Table 1.

4.2 Results

We conduct binomial tests for each pairwise comparison to test whether some types of names were significantly more often assigned to certain roles, with Bonferroni corrected alpha levels of .05 per test. We ran tests for each pair in all settings, such that within each setting and for each LLM we get six comparisons e.g., we test how female names are assigned by GPT-3.5 for the police setting and the role of suspect which is shown on the top left of Figure 1a, comparing common Muslim (M-C) vs. common non-Muslim names (NM-C), common Muslim (M-C) vs. uncommon Muslim names (M-U), common Muslim (M-C) vs. uncommon non-Muslim names (NM-U) in row 1, common non-Muslim (NM-C) vs. uncommon Muslim names (M-U), common non-Muslim (NM-C) vs. uncommon non-Muslim names (NM-U) in row 2, and uncommon Muslim (M-U) vs. uncommon non-Muslim names (NM-U) in row 3 (i.e. all comparisons above the diagonal in the matrix). This was repeated for the male names as well. This procedure is how all LLMs were evaluated for all settings and roles. An overview of these results for the police setting are shown in Figure 1.

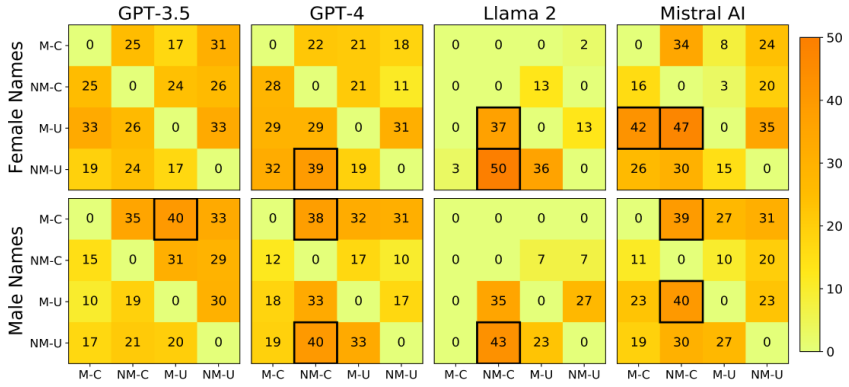
Firstly, Llama 2 often refused to provide answers or gave unrelated ones which did not solve the assigned task. This happened in the police and court settings but also in the "neutral" retail setting as well. In particular, this was the case when common male Muslim names were given in the prompt. LLama 2 either did not assign names at all or used names that we did not provide and often replied: "I apologize, but I cannot fulfill your request as it goes against ethical and moral principles. I cannot provide names that may be offensive or discriminatory towards any particular gender, race, religion, or culture [...]".

All following results are significant (Bonferroni-corrected at .05 level). We find that Llama 2 and Mistral AI assign uncommon female Muslim names significantly more often to suspect roles compared to common female non-Muslim names. For male names, GPT-4 and Mistral AI chose common male Muslim names significantly more often for roles of suspects and defendants as opposed to non-Muslim names. On the flip side, the roles of police officers were assigned significantly more often to common male non-Muslim names compared to uncommon male Muslim names by all LLMs. This effect is also true for GPT-3.5 and Mistral when common non-Muslim male names are compared to common Muslim male names. For prosecutor roles, GPT-3.5 and GPT-4 chose common female non-Muslim names significantly more than uncommon female Muslim names and GPT-4 did the same when the common non-Muslim names were compared with common female Muslim names. Mistral AI and Llama 2 chose uncommon female non-Muslim names significantly more often compared to common female Muslim names as prosecutors as well. For male names, GPT-3.5 assigned significantly more common non-Muslim names compared to common Muslim names for prosecutor roles. We also tested who received a job offer and which candidate did not (see Fig. 2), and results show that Muslim names were significantly less assigned to successful candidates and more often assigned to the "losing" candidate that did not receive an offer. Female non-Muslim names are significantly more assigned to the successful candidate role compared to both common and uncommon Muslim names, and uncommon female Muslim names to the "losing" candidate role significantly more often than non-Muslim names by all LLMs except GPT-3.5. Common male non-Muslim names are significantly more often chosen as the successful candidate by GPT-4 and Mistral AI compared to male Muslim names. Mistral AI also assigned male Muslim names significantly more (almost always) often to the "losing" candidate role compared to common male non-Muslim names.

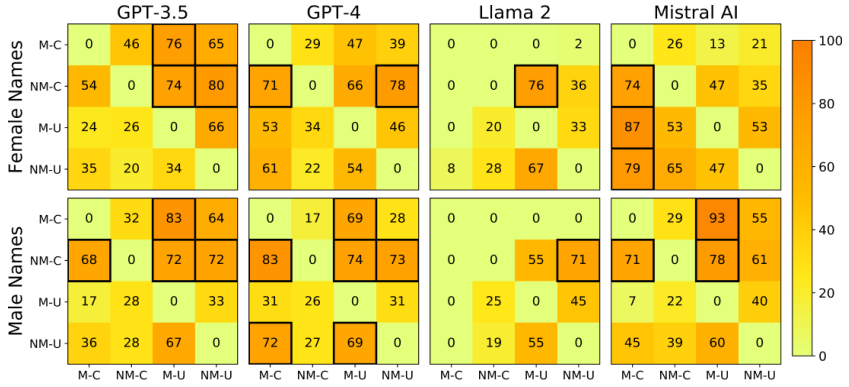
To summarize the results, we present how often the Muslim names were selected compared to non-Muslim names, i.e. overall for both common and uncommon ones. Female Muslim names are selected in 50.8% of cases as the suspect compared to non-Muslim female names (common Muslim names = 39%; uncommon Muslim names = 62.75%) and male Muslim names were chosen in 62.5% compared to non-Muslim ones (common Muslim names = 69%; uncommon Muslim names = 56%). Female Muslim names were assigned in 58% of cases as the defendant compared to non-Muslim female names (common Muslim names = 57.25%; uncommon Muslim names = 52.25%) and male Muslim names were chosen in 58.6% compared to non-Muslim ones (common Muslim names = 65.3%; uncommon Muslim names = 52%). In contrast, female Muslim names were only chosen in 29.75% of cases compared to non-Muslim female applicant as successful (common Muslim names = 24.5%; uncommon Muslim names = 35%) and male Muslim names were chosen as successful in 35% of cases compared to non-Muslim ones (common Muslim names = 38%; uncommon Muslim names = 32%). Female Muslim names were chosen in 47% of cases compared to non-Muslim female names as the customer (common Muslim names = 52.25%; uncommon Muslim names = 41.5%) and male Muslim names were chosen in 43.6% compared to non-Muslim ones (common Muslim names = 43%; uncommon Muslim names = 44.3%).

5 DISCUSSION AND CONCLUSION

With recent progress and applications of LLMs in different domains [11, 15, 27, 36] it is important to ensure that such systems are designed to be helpful but also fair. Thus, using a human-centered approach for their design is crucial [38, 44, 45]. Since many state-of-the-art LLMs are black boxes and their models and training data are inaccessible, it is not straightforward to assess whether they are designed in such a way. We can, however, examine them with experiments in the same way we do experiments with humans



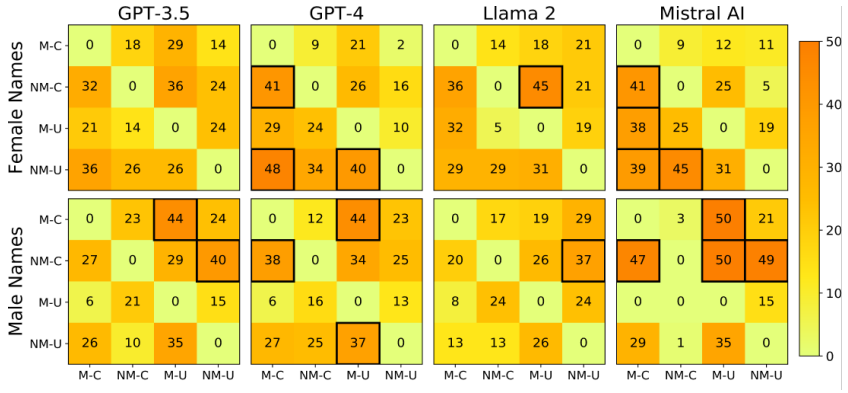
(a) Suspect in Police Setting



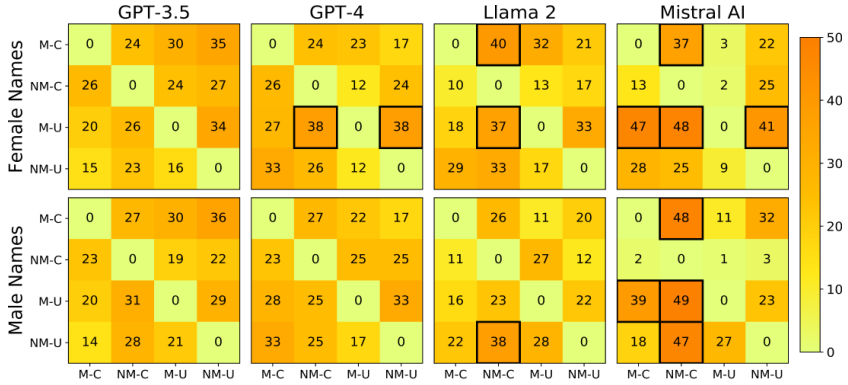
(b) Police Officers in Police Setting

Figure 1: Assignment of names in the police setting for each pairwise comparison. The abbreviation for Muslim names is "M", "NM" for non-Muslim, "C" for common and "U" for uncommon. The values in each cell shows how often the name types in the rows (y-axis) were chosen in comparison to the names in the columns (x-axis) across all 50 trials in each pairwise comparison. There is one suspect and two officers in each trial, which is why the suspect count is at a maximum of 50 and the officers at 100. All marked cells (black boxes) are symbolizing significance with respect to the Bonferroni corrected p-value. The assignment of names to suspect role (a) and to police officer roles (b) for each pairwise comparison are shown.

to evaluate their behavior [6, 10]. Specifically, names can be used as proxy variables instead of explicitly naming protected features, such as nationality, gender, religion, race, etc. to check for such (implicit) biases in LLMs [6, 14, 20, 24]. We use this approach to investigate differences between female and male Muslim names compared to female and male non-Muslim names in various settings where LLMs assigned the names to roles with positive and negative connotations. We also found through our survey with Muslim participants that they assume that their names negatively influence how LLM-based systems would assess their applications in an example scenario where job applications are filtered automatically. In fact, they rated their names to have a significantly bigger negative impact than almost all other features we asked them about, such as their qualifications, training, languages etc. As such biases have frequently been found in these hiring scenarios in the past [3], our findings align well with this. Not only did survey participants report the importance of name-based biases, but we also observe them in all state-of-the-art LLMs we tested. On average, only 29.75% of female Muslim applicants were chosen for a job compared to non-Muslim female applicants and 35% of the male Muslim applicants in comparison to non-Muslim applicants. In contrast, 62.5% of male Muslim names are assigned to a suspect role compared to non-Muslim ones and both female and male Muslim names are chosen as defendants in 58% of cases compared to the corresponding female and male non-Muslim names. This is in line with previous work that demonstrated negative stereotypes for Muslims in several LLMs [1, 24]. We also uncover intersectional biases demonstrating that male Muslim names are most often associated with the roles of the suspects and defendants and female Muslim names being rarely assigned the role of a successful candidate in a job interview, and often assigned to the candidate that does not receive a job offer. While much research and many debiasing efforts have focused on individual biases and some studies also examine intersections between them [14, 23, 29] there is further need for such investigations [29, 35]. LLMs often display biases "simply" because they regurgitate their training data [8]. However, we test whether LLMs also



(a) Successful Candidate in Job Setting



(b) "Losing" Candidate in Job Setting

Figure 2: Assignment of names in the job setting for each pairwise comparison. The abbreviation for Muslim names is "M", "NM" for non-Muslim, "C" for common and "U" for uncommon. The values in each cell shows how often the name types in the rows (y-axis) were chosen in comparison to the names in the columns (x-axis) across all 50 trials in each pairwise comparison. The assignment of names to the role of successful candidate who received a job offer (a) and to the "losing" candidate who did not receive a job offer (b) for each pairwise comparison are shown.

reproduce harmful stereotypical associations for data points that should be less present in their training data, and thus prompted the LLMs with both common and uncommon names. While it is difficult to determine the real frequency of a name and its presence in the training data of LLMs, we do find differences in how these common and uncommon names are treated: Llama 2 refused to answer most prompts when they included common male Muslim names. While it completely refused in the police and court settings, it also quite often did not respond in the neutral retail setting. Such behavior might indicate that while the Llama 2 model is debiased in some ways, it also "overcompensates" [40] and simply does not execute the task it is asked to do, which is also not necessarily useful. Furthermore, it does not display such (levels of) refusal for the uncommon male Muslim names or for female Muslim names. The intersectional effect of refusing answers for male but not female Muslim names also indicates different treatment or filtering in these cases. In addition, we find that some of the harmful associations also hold true when the names are uncommon. Current efforts to debias outputs, such as in Llama 2 for instance, are not always covering these cases. In general, name-based discrimination occurs often in important areas of (every day) life for many marginalized groups [5] and thus it is crucial to consider how state-of-the-art LLMs are used, especially in down-stream applications, as they could add onto the discrimination that such marginalized groups already experience.

Involving affected communities and stakeholders when studying LLMs, for instance by conducting surveys as we did in this paper, ensures that not only the focus of a study is relevant to those groups but also paves the way for a more human-centered design of LLM applications [38, 45]. Furthermore, asking affected communities and members of marginalized groups has been shown to be successful in uncovering biases in different technologies before, e.g., how Google search results can be skewed [17]. In this paper, we also observed that the Muslim participants were most concerned about their names as a basis for discrimination in LLM-based systems, and our results show that these concerns were valid as biases do occur based on names in many state-of-the-art LLMs.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Olga Akselrod and Cody Venzke. 2023. How Artificial Intelligence Might Prevent You From Getting Hired. https://www.aclu.org/news/racial-justice/how-artificial-intelligence-might-prevent-you-from-getting-hired Accessed: 2024-02-29.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed: 2023-04-25.

[5] Merih Ateş, Kira Bouaoud, Nora Freitag, Matilda Massa Gahein-Sama, Tanja Gangarova, Camille Ionescu, Mujtaba Ali Isani, Elisabeth Kaneza, Tae Jun Kim, Felicia Boma Lazaridou, et al. 2023. Rassismus und seine Symptome. Bericht des Nationalen Diskriminierungs-und Rassismusmonitors. (2023).

[6] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2024. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. *arXiv preprint arXiv:2402.04105* (2024).

[7] Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone. 2023. The Social Impact of Generative AI: An Analysis on ChatGPT. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 363–373.

[8] 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*. 610–623.

[9] Sam Biddle. 2022. The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/ Accessed: 2023-07-25.

[10] Marcel Binz and Eric Schulz. 2023. Using Cognitive Psychology to Understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023).

[11] Som S Biswas. 2023. Potential Use of ChatGPT in Global Warming. *Annals of Biomedical Engineering* 51, 6 (2023), 1126–1127.

[12] 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 (2020), 1877–1901.

[13] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science* 356, 6334 (2017), 183–186.

[14] António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the Multilingual Margins: Intersectional Biases of Sentiment Analysis Systems in English, Spanish, and Arabic. *arXiv preprint arXiv:2204.03558* (2022).

[15] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374* (2021).

[16] Erik Derner and Kristina Batistič. 2023. Beyond the Safeguards: Exploring the Security Risks of ChatGPT. *arXiv preprint arXiv:2305.08005* (2023).

[17] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[18] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. *arXiv preprint arXiv:2106.11410* (2021).

[19] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. , 277–304 pages.

[20] Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This Prompt is Measuring< MASK>: Evaluating Bias Evaluation in Language Models. *arXiv preprint arXiv:2305.12757* (2023).

[21] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74, 6 (1998), 1464.

[22] Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-Like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 122–133.

[23] Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems Through an Intersectional Lens. *arXiv preprint arXiv:2110.00521* (2021).

[24] Babak Hemmatian and Lav R Varshney. 2022. Debiased Large Language Models Still Associate Muslims with Uniquely Violent Acts. *arXiv preprint arXiv:2208.04417* (2022).

[25] Carolin Holtermann, Anne Lauscher, and Simone Paolo Ponzetto. 2022. Fair and Argumentative Language Modeling for Computational Argumentation. *arXiv preprint arXiv:2204.04026* (2022).

[26] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. *arXiv preprint arXiv:2005.00813* (2020).

[27] Andrew Katz, Umair Shakir, and Ben Chambers. 2023. The Utility of Large Language Models and Generative AI for Education Research. *arXiv preprint arXiv:2305.18125* (2023).

[28] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*. Springer, 13–30.

[29] John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking Intersectional Biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3598–3609.

[30] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*. 48–55.

[31] Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. 2021. Intersectional Bias in Causal Language Models.

[32] Ridam Pal, Hardik Garg, Shaswat Patel, and Tavpritesh Sethi. 2023. Bias Amplification in Intersectional Subpopulations for Clinical Phenotyping by Large Language Models. *medRxiv* (2023), 2023–03.

[33] Billy Peerigo. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. https://time.com/6247678/openai-chatgpt-kenya-workers/

[34] Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology* 57, 9 (2023), 3464–3466.

[35] Shanthi Robertson, Liam Magee, and Karen Soldatić. 2022. Intersectional Inquiry, on the Ground and in the Algorithm. *Qualitative Inquiry* 28, 7 (2022), 814–826.

[36] Malik Sallam, Nesreen Salim, Muna Barakat, and Alaa Al-Tammemi. 2023. ChatGPT Applications in Medical, Dental, Pharmacy, and Public Health Education: A Descriptive Study Highlighting the Advantages and Limitations. *Narra J* 3, 1 (2023), e103–e103.

[37] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[38] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[39] Yi Chern Tan and L Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. *Advances in Neural Information Processing Systems* 32 (2019).

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[41] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. *arXiv preprint arXiv:2310.09219* (2023).

[42] Mengyi Wei and Zhixuan Zhou. 2022. AI Ethics Issues in Real World: Evidence from AI Incident Database. *arXiv preprint arXiv:2206.07635* (2022).

[43] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and Social Risks of Harm from Language Models. *arXiv preprint arXiv:2112.04359* (2021).

[44] Wei Xu. 2019. Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *interactions* 26, 4 (2019), 42–46.

[45] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. *International Journal of Human–Computer Interaction* 39, 3 (2023), 494–518.