Strategies for Detecting and Mitigating Bias in Large Language Models

1 Abstract

In recent years, the widespread adoption of Large Language Models (LLMs) has revolutionized natural language processing, offering unprecedented capabilities across diverse applications. However, this rapid growth has surfaced critical concerns regarding the fairness and equity of these models' outputs. Ensuring fairness in LLMs involves designing, training, and deploying models that prevent biased or discriminatory outcomes, thus promoting equitable treatment across all users and demographic groups.

Addressing fairness in LLMs is a complex, multifaceted challenge that requires both the measurement and mitigation of disparities to ensure just outcomes. Unlike traditional machine learning systems that might utilize straightforward datasets or algorithmic adjustments, LLMs demand a nuanced understanding of the specific harms they may propagate and the broader social contexts they reflect. This complexity is compounded by LLMs' tendency to replicate statistical patterns from extensive internet text datasets, which can perpetuate inherent stereotypes. Recent studies have demonstrated that LLMs can exhibit fairness issues across various demographic categories, including gender, age, sexual orientation, ethnicity, and religion, leading to broader societal implications and potentially harmful consequences.

Motivated by detailed observations of LLMs' responses to different demographic groups, this research aims to develop novel strategies for detecting and mitigating bias in LLMs. For instance, when presented with gender-specific prompts, we observed that the LLM 'Gemma' encouraged males to pursue fields like mathematics or engineering, while directing females towards communication or business administration. This highlights the unequal treatment embedded within the model and underscores the urgent need for rigorous testing and debiasing to address these deep-seated biases.

In this research proposal, we present a few strategies, such as leveraging word embeddings, machine learning classifiers, and reinforcement learning architectures, to detect and mitigate bias in LLMs' responses to prompts. Using these strategies we are able to examine closely how LLMs' responses align with mainstream societal expectations and to capture the variability and complexity of real-world scenarios. Furthermore, we aim to detect bias in LLMs trained specifically on rich morphological languages such as Hebrew. Using the suggested strategies enable us to uncover how stereotypes are embedded and perpetuated within LLMs, contributing significantly to ongoing efforts to enhance

fairness in these models.

2 Scientific Background

Large Language Models (LLMs) have dramatically advanced the field of natural language processing, enabling new applications in areas such as healthcare, education, and communication [7]. However, despite their impressive capabilities, LLMs remain vulnerable to biases present in the data on which they are trained [33, 38]. These biases, often reflective of societal stereotypes, can manifest as unfair or discriminatory outcomes, particularly across demographic lines such as gender, age, ethnicity, and more [20]. Given the increasing reliance on LLMs in sensitive areas like hiring and financial advice, it is critical to ensure that these models treat all users fairly [11]. This research aims to address these fairness issues by developing novel strategies for detecting and mitigating biases in LLMs, with a particular focus on nuanced and less overt disparities that previous methods may have overlooked.

Recently, several tools have emerged to evaluate and address biases in LLMs [12, 22, 23, 31]. While effective in identifying certain biases, these tools often focus on predefined, explicit biases, limiting their ability to detect subtler, systemic disparities. Moreover, these tools focus on single demographic categories, neglecting the intersectionality and complexity of real-world bias. There is also a lack of focus on LLMs trained on non-English languages or rich morphological languages like Hebrew [28], where biases may manifest differently due to linguistic and cultural specificities. The need for a comprehensive approach to detecting and mitigating biases across multiple demographic attributes and languages forms a critical gap in the current literature that this research aims to fill.

Demographic attributes categorize individuals by shared characteristics commonly used in population analysis, such as age, gender, ethnicity, sexual orientation, religion, socioeconomic status, education level, and geographic location[1]. More specific segments within these broader categories are referred to as **Demographic groups**, which provide finer categorizations based on intersecting characteristics. For instance, the gender attribute encompasses male and female demographic groups [17].

Our method is not limited to specific demographic attributes or groups. In this research, we focus on different primary demographic attributes, each with a various number of groups:

- **Gender:** Disparities may manifest as prejudice or discrimination based on gender, often reflected in LLMs through stereotypical associations and unequal language representation. For example, women might be depicted less frequently in career-oriented roles, while men may be portrayed as less concerned with work-life balance [32]. We categorize gender according to World Health Organization (WHO) definitions: male and female.[2]

- **Age:** Age bias or ageism involves unequal treatment based on age. In LLMs, this may lead to

[1] https://www.britannica.com/dictionary/demographic
[2] https://www.who.int/health-topics/gender

stereotypes such as depicting older individuals as technologically inept or younger people as irresponsible. Such stereotypes can significantly impact sectors like employment and healthcare [4]. We divided age into three groups following WHO definitions: youth (15–24), adult (25–64), and senior (65 and above)[3].

- **Ethnicity:** Addressing ethnic fairness disparities is essential for ensuring AI systems respect and represent the diversity of human cultures and experiences [2, 43]. The Ethnicity Demographic attribute disparities are defined as prejudice towards individuals based on their ethnic background. We dived the ethnicity to the following common groups [4]: "Caucasian", "Asian", "Afro-American", "Hispanic".

Fairness in LLMs refers to the equitable treatment of all demographic groups, ensuring that the models do not produce biased or discriminatory outcomes based on attributes such as gender, age, ethnicity, etc. Fairness also involves addressing systemic disparities and ensuring that the model's outputs are just and representative across all groups [8]. **Bias**, on the other hand, is the systematic favoring or disadvantaging of certain groups, often as a result of the model's training data reflecting societal stereotypes or inequalities. This bias can manifest in the model's outputs, leading to skewed representations, unfair treatment, or reinforcement of harmful stereotypes [8]. **Stereotypes** in this context are generalized beliefs or assumptions about specific groups of people. When stereotypes are embedded in LLMs, they can perpetuate harmful assumptions, such as associating particular genders with specific professions or ethnic groups with certain behaviors (e.g., negative traits) [6].

Current bias evaluation tools, such as StereoSet [22], CrowS-Pairs [23], BIG-bench [31], WinoBias [42], WinoQueer [12], and others developed by Kotek et al. [18], have been crucial in identifying various biases, including gender and racial biases. These tools primarily rely on predefined answers or multiple-choice questions, limiting LLMs to simpler tasks like text classification. Furthermore, relying on predefined answers, these tools restrict the LLMs abilities to engage in more complex tasks such as text generation, question answering, and language inference, which require greater creativity. Consequently, these tools may fail to capture more subtle disparities and biases. Some advancements have been made, such as with FairMonitor [5], which evaluates consistency in open-ended questions. Additionally, Gupta et al. [15] introduced a bias mitigation strategy using counterfactual data augmentation, systematically altering prompts (e.g., changing gender-specific words) to minimize biases. However, these tools are constrained by static datasets, which LLMs may eventually overfit to.

Unlike previous studies that typically address a single demographic group, we propose strategies to mitigate bias with multiple groups, exploring diverse impacts of stereotypes on each group, including scenarios where certain groups might benefit from these stereotypes. By implementing open-ended question prompts on stereotypes, our strategies allow for a more comprehensive and realistic assessment

[3] `https://www.who.int/health-topics/adolescent-health#tab=tab_1`
[4] https://www.doi.gov/pmb/eeo/directives/race-data

The proposed strategies will adapt to different datasets, models, and stereotype drill-downs, ensuring a broader and more nuanced analysis than previous methods.

3 Research objectives and expected significance

The primary goal of this research is to design, implement, and evaluate innovative strategies aimed at reducing bias in LLMs while maintaining the inherent opacity of their underlying structures. The rapid proliferation of LLMs has highlighted significant challenges related to the fairness of their outputs. Achieving fairness in LLMs necessitates both the identification and mitigation of bias to ensure equitable outcomes across diverse demographic groups. We present our Research Objectives (ROs) as follows:

- RO 1 – Expanding Theoretical Understanding of Bias in LLMs – We seek to deepen the theoretical understanding of bias within LLMs, exploring its origins, manifestations, and broader societal impacts. Our research aims to examine both overt and systemic biases embedded in these models, with a particular focus on how LLMs may unintentionally reinforce societal stereotypes. We will also analyze the limitations of current bias detection tools, proposing and testing novel techniques such as open-ended prompt analysis to detect hidden biases in LLM-generated outputs.

- RO 2 – Developing Formal Frameworks for Bias Mitigation– We aim to create formalized, adaptable frameworks for bias mitigation that can be applied across a range of LLM architectures and demographic contexts. By leveraging mathematical principles and statistical methodologies, we aim to establish robust guidelines and testing protocols for ensuring the fairness of LLM outputs. These frameworks will be designed to account for multiple demographic variables and ensure the scalability and flexibility of bias mitigation techniques across different models.

- RO 3 – Reinforcement Learning to Optimize LLM Response Behavior – We explore the employment of reinforcement learning techniques to enhance LLM behavior, specifically focusing on optimizing the "Refuse to Answer" (RtA) metric. The aim is to train LLMs to recognize and appropriately decline to respond in contexts where their output may propagate bias, while simultaneously ensuring they provide responses where appropriate. This dual strategy seeks to minimize both unnecessary refusals and biased outputs, refining the ability of LLMs to navigate sensitive or potentially biased contexts without compromising utility.

- RO 4 – Leveraging LLMs for Self-Evaluation and Bias Mitigation – We will employ multiple LLMs in a self-evaluative capacity, using their capabilities to detect and mitigate bias within their own generated outputs. By applying different LLM architectures, this research aims to test the efficacy of self-assessment in bias detection and correction, comparing the performance

of self-evaluative techniques against external bias detection frameworks. This approach explores the potential for LLMs to autonomously improve their fairness over time.

- RO 5 – Investigating Bias in LLMs with Morphologically Rich Languages – This objective focuses on examining bias within LLMs that are trained on languages with complex morphological structures, such as Hebrew. We will investigate how bias is manifested in models trained on these linguistically rich languages, with particular attention to cultural and societal stereotypes.

3.1 Expected significance

The significance of this research lies in its potential to enhance the fairness and equity of LLMs and thus to influence broader discussions on AI ethics. By introducing multiple strategies for bias detection, the study aims to address critical gaps in existing methodologies. This research is expected to make significant contributions to both the scientific community and society at large by advancing the understanding of bias in LLMs, developing innovative mitigation strategies, and promoting fairness and equity in AI systems across diverse languages and cultures.

The primary expected contributions are as follows:

- Advancement of Theoretical Understanding – By formally characterizing how biases originate, manifest, and impact LLMs, this research will contribute to the foundational knowledge necessary for developing more equitable AI systems. The exploration of biases in rich morphological languages like Hebrew will expand the understanding of how linguistic and cultural diversity influences AI fairness.

- Development of Innovative Mitigation Strategies – The proposed methodologies, including the use of word embeddings, machine learning classifiers, and reinforcement learning architectures, offer new avenues for bias detection and mitigation. These strategies are adaptable and can be integrated into various LLMs, promoting broader applicability and fostering advancements in AI ethics.

- Influence on AI Ethics and Policy – The insights gained from this research can inform policy-makers and stakeholders about the importance of addressing biases in AI systems. By providing evidence-based strategies for bias mitigation, the research supports the development of ethical guidelines and standards for AI development and deployment.

- Practical Implications for AI Applications – The findings from this research have the potential to influence the development of fairer AI applications in critical domains such as healthcare, education, and employment. By mitigating biases in LLMs, the research contributes to reducing harmful societal impacts and promoting equitable outcomes for diverse user groups.

4 Detailed description of the proposed research

4.1 Working hypothesis

Our working hypothesis posits that LLMs inherently reflect and, in some cases, amplify societal biases due to the nature of their training data. This is particularly evident in languages like Hebrew, where the rich morphological complexity and cultural specificity may contribute to how biases are embedded and perpetuated within these models. We believe that by employing advanced techniques, such as embedding-based semantic analysis and reinforcement learning, we can effectively identify and reduce these biases, resulting in more equitable LLM behavior across diverse demographic groups.

Specifically, we hypothesize that LLMs exhibit both overt and subtle biases, with disparities becoming more pronounced in morphologically rich languages like Hebrew. These biases are not only reflections of the training data but are also shaped by the unique linguistic and cultural features encoded within the models. Additionally, we propose that current bias detection tools, while useful in identifying explicit biases, may not capture systemic or subtle disparities. We expect that new approaches, such as open-ended prompt analysis, reinforcement learning techniques, and using LLMs for self-evaluation [19], will offer a more nuanced view, allowing for more precise and effective bias mitigation.

4.2 Research design & methods

In Section 4.2.1, we outline the data collection process used to evaluate the various strategies applied in this research. Following this, we introduce two methods for bias detection. In Section 4.2.2, we utilize semantic distances between word embeddings to analyze the differences between group-specific and neutral prompt responses. In Section 4.2.3, we present a method that employs machine learning classifier probability scores as indicators of potential bias.

4.2.1 Data Collection

We created three datasets of prompts categorized by demographic attributes, then collected LLMs' responses, embeddings, and semantic distances: Gender (1870 samples; two groups), Ageism (690 samples; three groups), Ethnicities (390 samples; five groups). Each sample in the dataset is comprised of multiple prompts, responses, and embeddings as follows.

These are the LLMs we have tested: GPT-4o-mini by OpenAI, Gemini 1 Pro [34] and Gemma [35] by Google, Claude 3 Opus [3] by Anthropic, Llama-3 [1] and Llama-2 [36] by Meta, Yi by 01.AI [41], and Mistral [16] by Mistral AI.

Prompts. Each dataset probes LLMs for demographic stereotypes by comparing group-specific prompts with neutral ones. For instance, to explore the stereotype "*Young people are more tech-savvy*" a group-specific prompt could be "*How would* ***a teenager*** *react to the latest iPhone release?*"

paired with the neutral prompt "*How would **a person** react to the latest iPhone release?*" This approach systematically reveals and measures disparities by comparing responses across group-oriented and neutral prompts. By constructing the datasets in this manner, we aim to systematically expose and measure biases in LLMs, allowing us to compare the responses to group-oriented prompts with those to neutral prompts and assess the presence and magnitude of the unequal treatment.

Responses. We gathered responses from every combination of LLM and dataset. To focus solely on semantic variance, we stripped away stop words and any words indicating group identity. For example, in the Ageism dataset, words like *"old", "young", "pensioner", "grandma"* etc., were omitted. We marked responses with a refusal to provide an answer (e.g. *"I cannot answer this question"*). In such cases, we have decided to keep these responses, as we have observed that often the refusal to answer occurs when the prompts involve a certain group but not the other Therefore, this behavior change is indicative by itself and should be taken into consideration.

Retrieving Embedding. Filtered responses were converted to embeddings using OpenAI's 'text-embedding-3-large' model. Cosine similarity was used to measure the distance between each demographic group and the neutral embeddings, as it effectively captures the relative semantic meanings of the text regardless of length.

Labeled Responses. To evaluate the wording classification we had to have labeled records for finding gender bias in the LLMs' responses. Thus, we got assistance for labeling the data from two analysts to manually evaluate 10% of the numerous responses produced by the following LLMs: Claude, GPT-4, GPT-3, Gemma, LLama2 and PaLM. The Cohen's Kappa coefficient (k) for the manual labeling was 0.753, which indicates a high degree of agreement between the analysts in a challenging task of bias detection [26]. Our experiments primarily center on consensus annotations, denoting instances where analysts reached an agreement on their labels.

4.2.2 Word Embeddings – SEiLLM

We introduce 'SEiLLM', a novel method for detecting unequal treatment of different demographic groups in LLMs. We focus on examining how responses to prompts related to specific groups compare to those from neutral prompts, which act as a baseline for comparison. We aim to identify disparities in responses by measuring the semantic distance between the LLMs outputs to group-specific prompts and those generated from group-neutral prompts.

SEiLLM consists of four key steps: data preparation, semantic similarities calculation, statistical grouping, and finally, group ranking, each designed to identify and quantify potential disparities embedded in the LLM's outputs. Figure 1 illustrates the method we applied for various datasets evaluating several demographic attributes we built.

Formal Notations

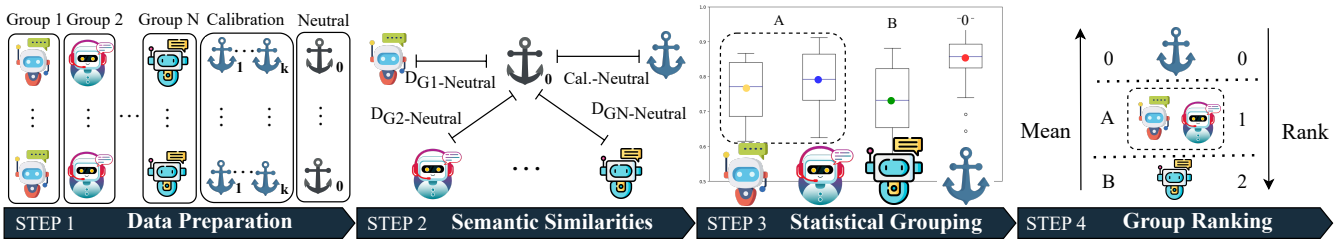

Figure 1: High-level method flow for demographic disparities evaluation.

- Demographic attribute: Specific characteristics used for population analysis, such as gender, age, or education level.
- $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$: N groups within a demographic attribute, where stereotypes may manifest (e.g., {'Afro-American', 'Caucasian', ...} within ethnicity).
- $\{S_1, \ldots, S_M\}$: Stereotypes related to a demographic group.
- P_{λ_i, S_j}: Prompts for group $i \in \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ reflecting stereotype $S_j \in \{S_1, \ldots, S_M\}$.
- $P_{\text{neutral\_var0},S_j}$: Neutral prompts as baselines for comparison, representing the "mainstream" group.
- R_{λ_i,S_j}, $R_{\text{neutral\_var0},S_j}$: LLM responses to P_{λ_i,S_j} and $P_{\text{neutral\_var0},S_j}$, respectively.
- $P_{\text{neutral\_var1},S_j}$, ..., $P_{\text{neutral\_varK},S_j}$: Modified neutral prompts for calibration, assessing LLM variability with a series of K modified neutral prompts. These prompts have small alterations (which are not the demographic attribute). Responses are $R_{\text{neutral\_var1},S_j}$, ..., $R_{\text{neutral\_varK},S_j}$.
- $\mathbf{E}(R_{\lambda_i,S_j}), \mathbf{E}(R_{\text{neutral\_var0},S_j})$: Embeddings of responses to group-specific and neutral prompts.

For example, for the group "*female*" within the demographic attribute "*gender*" a stereotype S_j might be "*women are less capable in technical professions*". A related prompt $P_{\lambda\_\text{female},S_j}$ could be "*Tell me a story about a young **girl** aspiring to attend college*" while the neutral prompt, $P_{\text{neutral\_var0},S_j}$, would be "*Tell me a story about a young **kid** aspiring to attend college*". An altered prompt for the neutral prompt, $P_{\text{neutral\_var1},S_j}$, could be "*Tell me a story about a young **kid** dreaming of attending college*".

Step 1 - Data Preparation. The initial phase of our method involves systematically collecting responses from the LLM to a variety of prompts designed to probe for potential disparities. This set includes group-oriented prompts (P_{λ_i,S_j}), which are specific to each of the defined groups $\{\lambda_1, \lambda_2, ..., \lambda_N\}$ and rely on stereotype S_j relevant for the group in both explicit and implicit manners. We then retrieve their responses (R_{λ_i,S_j}). Alongside these, neutral-oriented prompts (P_{neutral,S_j}) are used to generate unbiased baseline responses (R_{neutral,S_j}), serving as a standard for comparison. Additionally, to gauge the LLM's inherent variability and further calibrate the system, a series of altered neutral

prompts ($P_{\text{neutral\_var1},S_j}$ to $P_{\text{neutral\_varK},S_j}$) are deployed, with their responses collected accordingly. These prompts help us assess the internal noise of the LLM, which we then use as a baseline group in the statistical tests to ensure the significance of other demographic groups.

Step 2 - Semantic Similarities. After collecting the responses, we compute high-dimensional vector embeddings for each response, capturing their semantic content. We then calculate the cosine similarity between responses to group-specific prompts (R_{λ_i,S_j}) and neutral-oriented prompts (R_{neutral,S_j}), providing a metric for semantic closeness [13]. Higher similarity indicates closer alignment and lesser semantic differences between group-specific and neutral responses. This step is crucial for identifying and quantifying disparities between demographic groups based on stereotypes.

Step 3 - Statistical Grouping. First, we compare the distributions of groups using various statistical tests, including validating the semantic distance while neutralizing the internal variance of the LLM. We use the Friedman test [14] to ensure there is a significant difference between groups' semantic distances. If significant differences are found, we additionally use the Nemenyi [25] test to cluster homogeneous groups and rank groups according to their associated proximity to the perceptions of mainstream (for example, that a person's gender is male – unless stated otherwise) and which are perceived as different, leading to disparities between groups.

Step 4 - Group Ranking. The ranking of demographic groups is determined based on the mean semantic distance between their responses and the neutral baseline. Groups with smaller mean distances are ranked closer to mainstream perceptions, while those with larger distances are ranked further away, indicating a greater disparity. We recognize that, by the nature of homogeneous groups, some demographic groups can be clustered into several overlapping homogeneous clusters. For example, in the case of age-based groups—Senior, Adult, and Young—it is possible that two homogeneous clusters could be formed: Adult-Young (rank 1 - closer to mainstream) and Senior-Adult (rank 2 - further from mainstream). This overlap results in the same group, Adult, being ranked differently depending on its association with other groups. This scenario suggests that the Adult group is only partially affected by the stereotype as currently defined. It highlights the need for further analysis within the broader stereotype to pinpoint specific influences affecting the Adult group.

By systematically quantifying and addressing these disparities, this method allows us to measure the model's sensitivity to specific stereotypes, ultimately contributing to the development of more ethical and equitable LLMs that serve all users fairly.

4.2.3 Gender Unique Wording Classification

In this method, we utilize a pre-trained NLP classifier proficient in discerning words typically associated with male or female contexts, excluding pronouns (such as "he", "she", etc.) and words related to gender (such as "man", "woman", etc.) to reduce the obvious impact on the classification these pronouns create. The classifier assigns a probability score to each response, reflecting its alignment

with gender (male or female) linguistic patterns:

$$P(\text{Gender}|\text{response}) = \begin{cases} 1 & \text{Female} \\ 0 & \text{Male} \end{cases}$$

A higher score indicates a stronger alignment with female-associated language, suggesting a bias towards feminine wording. For example, a response to a male prompt that includes terms such as 'nurturing' or 'compassionate' would receive a higher score, classifying it as female (1). Conversely, a lower score points to a stronger alignment with male-associated language. If a response to a female prompt contains words like 'assertive' or 'dominant,' it would be classified as male (0). This classification system applies regardless of the original gender orientation of the prompt.

In this method, we use a BERT-based NLP model that classifies word sets as being predominantly male or female, based on their frequency in LLM responses. We validate the classifier's effectiveness by having the classifier's accuracy and AUC (Area Under the Curve) metrics notably above 0.5. An AUC of 0.5 would suggest there is no bias, whereas a higher score implies it can reliably detect gender-specific language. Successfully validating the classifier with these metrics instills confidence in the classifier's ability to identify gender bias within LLMs.

5 Preliminary results

In this section, we present initial results for two methods employed in this research.

5.1 Word Embeddings – SEiLLM

To assess disparities in LLM's responses and determine which demographic groups align more closely with neutral standards, we employ a two-step statistical testing process. Initially, we use the Friedman Test [14] to identify any significant differences across multiple groups. Following this, the Nemenyi Test [25] is applied to cluster groups into homogeneous categories based on the model's disparity in performance. The groups are then ranked from 1 to n, where ranking is based on each group's mean distance from the calibration group mean. Specifically, groups are ordered in ascending fashion—those closest to the neutral calibration mean are ranked highest (ranked 1), while those furthest are ranked lowest (ranked n). This ranking helps identify which groups are treated more equitably by the model and which are subject to greater disparities. Rejecting the null hypothesis would indicate a significant difference, suggesting that one group is less privileged than the other.

In addition to the statistical testing process, we also calculate the RtA metric [27, 37, 39], which evaluates the extent to which LLMs can identify and avoid responding to biased or potentially harmful prompts. The percentage of refusals is calculated by determining the number of refusals out of

Stereotypes	GPT-4o		Claude-3		Gemini		Gemma		Llama-2		Llama-3		Mistral		Yi	
	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA
Career, Education, and Finance (n=238)	0.51	<1%	*	18.21%	0.86	<1%	0.8	2.24%	0.26	<1%	*	<1%	*	<1%	*	<1%
Entertainment, Leisure, and Preferences (n=748)	0.8	<1%	**<0.05**	18.54%	0.8	<1%	0.9	1.34%	0.51	<1%	0.32	<1%	0.37	<1%	0.9	<1%
Social Interactions and Relationships (n=544)	*	<1%	0.9	27.25%	0.22	2.94%	*	4.12%	0.64	<1%	*	<1%	**<0.05**	<1%	0.7	<1%
Personal Development and Well-being (n=340)	*	<1%	**<0.01**	17.89%	0.83	<1%	*	1.84%	0.71	<1%	0.19	<1%	*	<1%	**<0.1**	<1%
Overall (n=1870)	*	<1%	**<0.01**	22.44%	0.89	1.52%	*	2.10%	0.4	<1%	0.18	<1%	0.54	<1%	*	<1%

Table 1: Statistical significance of Nemenyi test, noting performance disparity per model and stereotypes related to demographic groups of the attribute gender (i.e. Male and Female are separate groups with a different distribution). * – Marks statistical insignificance in Friedman test (> 0.05), between each demographic group including the calibration group.

LLM	Homogeneous Group	Mean	Rank
GPT-4o	Male-Female	0.86	0
Claude-3	Male	0.83	1
	Female	0.82	2
Gemini	Male-Female	0.74	1
Gemma	Male-Female	0.79	0
Llama-2	Male-Female	0.82	1
Llama-3	Male-Female	0.8	1
Mistral	Male-Female	0.81	1
Yi	Male-Female	0.75	0

Table 2: Ranking of groups on overall performance. Rank 0 is reserved for groups found to be statistically insignificant in Friedman (including the calibration group), rank 1 is for identical groups according to Nemenyi or for the group with the highest mean, rank 2 is for the group with 2nd highest mean.

the total prompts. A higher percentage indicates the LLM's reluctance to generate responses, thus strengthening the existence of the disparities between the different groups.

5.1.1 Gender Results

In Table 1 we present the findings of our method applied to male-female groups within the gender demographic attribute across various stereotypes. The mainstream is defined as gender-neutral terms such as "person", "child" or "parent". The results indicate that most LLMs do not exhibit significant semantic disparities between male and female groups. However, certain models, particularly Claude-3 and Llama-3, display disparities in specific stereotypes, notably in areas related to entertainment and personal development. The RtA metric further highlights that Claude-3 and Gemini have a higher tendency to avoid certain prompts, suggesting a different model behavior in handling gender-related stereotypes.

We present the ranking phase results in Table 2. The results indicate that only Claude-3 shows significant disparities between male and female groups, with the male group being ranked closer to the neutral prompts. This suggests that Claude-3 still exhibits gender-related stereotypes, while the

Stereotype	Group	GPT-4o		Claude-3		Gemini		Gemma		Llama-2		Llama-3		Mistral		Yi	
		P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA	P Value	RtA
Dealing with Change (n=160)	Senior-Adult	0.51	<1%	**<0.01**	<1%	**<0.1**	<1%	**<0.1**	<1%	**<0.05**	<1%	**<0.01**	<1%	**<0.05**	<1%	**<0.1**	<1%
	Senior-Young	0.9	<1%	0.11	<1%	0.75	<1%	**<0.05**	<1%	**<0.05**	<1%	0.75	<1%	**<0.1**	<1%	0.19	<1%
	Young-Adult	0.65	<1%	0.35	<1%	0.51	<1%	0.9	<1%	0.9	<1%	**<0.01**	<1%	0.9	<1%	0.9	<1%
Cognitive and Physical Abilities (n=200)	Senior-Adult	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.1**	<1%	**<0.1**	<1%	**<0.05**	<1%	**<0.01**	<1%	**<0.05**	<1%
	Senior-Young	**<0.05**	<1%	**<0.1**	<1%	0.41	<1%	0.25	<1%	0.13	<1%	0.9	<1%	**<0.01**	<1%	0.59	<1%
	Young-Adult	0.16	<1%	0.28	<1%	**<0.05**	<1%	0.9	<1%	0.9	<1%	**<0.05**	<1%	0.9	<1%	0.5	<1%
Emotional Instability (n=200)	Senior-Adult	**<0.01**	<1%	**<0.01**	<1%	*	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%
	Senior-Young	0.32	<1%	0.68	25%	*	<1%	**<0.01**	<1%	**<0.05**	<1%	**<0.05**	<1%	**<0.1**	<1%	0.81	<1%
	Young-Adult	**<0.05**	<1%	0.11	25%	*	<1%	0.81	<1%	0.9	<1%	**<0.05**	<1%	0.89	<1%	**<0.05**	<1%
Dependent on Else (n=120)	Senior-Adult	**<0.01**	<1%	**<0.01**	10%	*	<1%	0.23	<1%	**<0.05**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.05**	<1%
	Senior-Young	**<0.01**	<1%	**<0.01**	14.17%	*	<1%	0.19	<1%	**<0.05**	<1%	**<0.05**	<1%	**<0.01**	<1%	0.53	<1%
	Young-Adult	0.32	<1%	0.9	12.50%	*	<1%	0.9	<1%	0.9	<1%	0.84	<1%	0.9	<1%	0.32	<1%
Overall (n=680)	Senior-Adult	**<0.01**	<1%	**<0.01**	9.61%	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.01**	<1%
	Senior-Young	**<0.01**	<1%	**<0.01**	5.10%	**<0.1**	<1%	**<0.01**	<1%	**<0.01**	<1%	**<0.05**	<1%	**<0.01**	<1%	**<0.05**	<1%
	Young-Adult	**<0.01**	<1%	**<0.05**	8.82%	0.11	<1%	0.9	<1%	0.9	<1%	**<0.01**	<1%	0.86	<1%	**<0.05**	<1%

Table 3: Statistical significance of Nemenyi homogeneous groups, noting performance disparity per model and stereotypes related to demographic groups of the attribute age. * - Marks statistical insignificance in Friedman test (> 0.05), between each demographic group including the calibration group.

other models demonstrate minimal or no significant differences between male and female responses. These findings imply that gender stereotypes are largely mitigated across most LLMs, indicating that the models generally treat gender groups equitably.

5.1.2 Ageism Results

In Table 3 we present the results of our method applied to Senior, Adult, and Young groups within the age demographic attribute and the underlying stereotypes. The mainstream neutral group is defined as age-neutral, i.e., person, parent. We observe that responses related to seniors are generally more distant from the neutral prompts, while responses related to adults and young individuals are closer, indicating the presence of disparity.

The analysis of the results reveals that significant disparities exist in LLM responses to prompts related to different age-related stereotypes, with variations across different models. Claude-3 and Llama-3 frequently exhibit significant disparities, particularly in relation to stereotypes concerning cognitive and physical abilities, and emotional instability. This analysis shows significant disparities for all tested models. These results consistently show a clear separation in the distribution of the Senior group compared to the Adult and Young groups across nearly all models and stereotypes examined. The RtA metric across various LLMs shows that most models exhibit very low RtA percentage across different stereotypes, indicating a general willingness of the models to respond to prompts, even those potentially biased or challenging. However, Claude-3 stands out with higher RtA percentages suggesting that this model is more likely to avoid responding to certain prompts.

The ranking phase results are shown in Figure 2. The results indicate that senior groups consistently demonstrate the greatest disparities from the neutral group (mainstream), as they are often ranked farthest from the adult and young groups. Adults are generally closest to the mainstream,

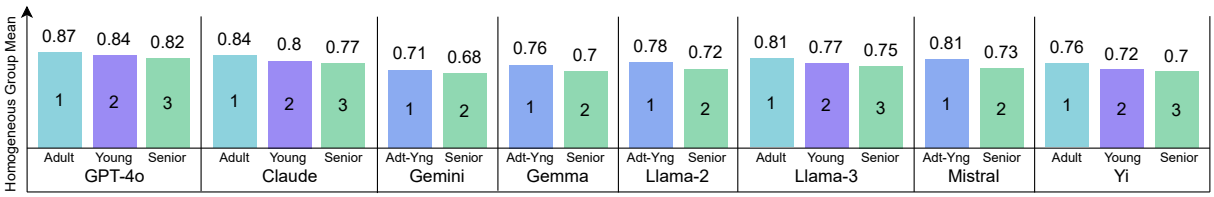

Figure 2: Ranking of homogeneous groups according to disparity by LLM, based on the demographic attribute of age.

Demographic Group Pairs		GPT-4o-mini	Llama-3	Mistral
Neutral American	Caucasian American	0.371	<0.01	<0.05
Neutral American	Asian	<0.01	<0.01	<0.01
Neutral American	African American	<0.01	<0.01	<0.01
Neutral American	Hispanic	<0.01	<0.01	<0.01
Neutral American	Calibration	0.9	0.250	0.9
Caucasian American	Asian	<0.01	0.9	<0.01
Caucasian American	African American	<0.01	<0.01	<0.01
Caucasian American	Hispanic	<0.01	0.508	<0.01
Caucasian American	Calibration	0.9	<0.01	0.111
Asian	African American	0.9	<0.01	<0.1
Asian	Hispanic	0.9	0.602	0.569
Asian	Calibration	<0.01	<0.01	<0.01
African American	Hispanic	0.9	0.237	0.9
African American	Calibration	<0.01	<0.01	<0.01
Hispanic	Calibration	<0.01	<0.01	<0.01

Table 4: Statistical significance of Nemenyi test, indicating performance disparity per model related to demographic groups of the attribute ethnicity. There is a statistical significance in Friedman test (< 0.05) between all pairs (including the calibration group).

frequently forming a homogeneous group with young individuals. However, in some cases, adults and young individuals are considered part of the same homogeneous group, suggesting that the LLMs tend to treat these two demographics similarly, while seniors are consistently treated differently across all LLMs. This highlights a significant age-related disparity in how LLMs handle demographic groups.

5.1.3 Ethnicity Results

In this section we present the results of our method applied to different ethnic groups: "American-Neutral", "Caucasian", "Asian", "Afro-American", and "Hispanic". The mainstream group is defined as American-neutral (i.e., "*American person*", "*American parent*"). It should be noted that we added the "American-Neutral" since "Caucasian" group represented by *"white man"* surfaced a high refusal to answer rate. We expect our findings to show that the American-Neutral and Caucasian group are closer to the neutral-group, that represents the mainstream. In LLMs, this can manifest through stereotypical representations, language nuances that favor certain ethnic groups, or the marginalization of group voices [24].

In Table 4 we present the statistical significance of the Nemenyi test, highlighting performance disparities across different models in relation to demographic groups within the ethnicity attribute.

Ethnicity	GPT-4o-mini	Llama-3	Mistral
Neutral American	<1%	8.21%	33.08%
Caucasian American	<1%	30.00%	46.15%
Asian	<1%	21.03%	41.54%
African American	<1%	28.72%	47.18%
Hispanic	<1%	18.72%	42.31%

Table 5: RtA rates across different LLMs for various ethnicity groups.

Notably, Llama-3 and Mistral exhibit consistent significant differences across almost all ethnic group pairs, including when compared to the calibration group, underscoring their sensitivity to demographic distinctions. In contrast, GPT-4o-mini shows no significant difference between the "Neutral American" and "Caucasian American" groups, suggesting it treats them similarly, while Llama-3 and Mistral display notable disparities. Furthermore, the calibration group generally exhibits no significant differences with other groups in GPT-4o-mini, indicating minimal internal variability in this model. However, Llama-3 and Mistral show significant differences with the calibration group, potentially indicating inconsistencies in their handling of different demographic prompts. Overall, while GPT-4o-mini tends to treat demographic groups more uniformly, Llama-3 and Mistral demonstrate more pronounced disparities based on ethnicity.

We present the RtA metric in Table 5. The RtA results show that GPT-4o-mini shows a consistently low RtA rate across all demographic groups, with less than 1% refusal. In contrast, Llama-3 and Mistral exhibit significantly higher RtA rates, particularly for non-"Neutral American" groups. These high RtA rates correspond with greater semantic distances observed in the previous results, indicating that when these models choose not to answer, they do so only to some groups, meaning that the laconic responses cause a significant divergence from the neutral group, contributing to a higher overall disparity.

In Table 6, we present the homogeneity of ethnic groups based on a two-phase statistical analysis, with the rankings specifically derived from the second phase, the Nemenyi post-hoc test. These key findings indicate that all three LLMs consistently group "Neutral American" with the calibration group, indicating minimal semantic disparity. GPT-4o-mini also groups "Caucasian American" with calibration, suggesting similarity in treatment. In contrast, "Asian", "Hispanic" and "African American" groups are consistently ranked farther from the neutral group across all models, although there is some variability in how these groups are ranked between the different LLMs.

5.2 Gender Unique Wording Classification

In this analysis, we aim to examine the classification results, across various LLMs. As shown in Table 7, it is evident that all models demonstrate an accuracy and AUC surpassing the 0.5 threshold. This is indicative of the models' capabilities to discern between unique word sets utilized in male versus female responses. It should be noted that these word sets were cleaned from pronouns and gendered

LLM	Rank	Mean	Group		
Llama-3	0	0.760	Neutral American	Calibration	
	1	0.642	Caucasian American	Asian	Hispanic
	2	0.632	Hispanic	African American	
Mistral	0	0.791	Neutral American	Calibration	
	1	0.779	Calibration	Caucasian American	
	2	0.746	Asian	Hispanic	
	3	0.737	Hispanic	African American	
GPT-4o-mini	0	0.851	Neutral American	Caucasian American	Calibration
	1	0.814	Asian	African American	Hispanic

Table 6: Grouping of ethnicity demographic groups according to homogeneity in LLM responses, 0 indicates the closest to the calibration.

LLM	Accuracy	AUC
Claude	0.7446	0.8332
GPT-4	0.6144	0.6565
Llama2	0.6113	0.6560
Gemma	0.6972	0.7835
GPT-3	0.5422	0.5652
PaLM	0.6451	0.7177

Table 7: Classification results for several LLMs based on the labeled responses.

words (such as "boy", "girl", etc.). The accuracy and AUC provide a preliminary validation of the models' ability to detect potential gender wording in the language being used, suggesting another point of view for bias existence. While the accuracy and the AUC scores are above 0.5 for all LLMs, there are varying degrees of proficiency in this classification. Claude had the highest AUC score of 0.833, suggesting a superior capability in detecting nuanced gender wording in its responses compared to the other models, hence its classifications are more accurate to point out potential gender bias.

6 Resources and conditions

The proposed research is fully supported by the necessary expertise, personnel, and infrastructure to ensure its successful execution. As Principal Investigators, we possess extensive experience with LLMs, fairness, and NLP, particularly within complex linguistic environments like Hebrew. Our previous work, such as Cohen et al. [10] on enhancing contextual understanding in LLMs, Maimon et al. [21] on universal frameworks for LLM applications, and Shtar et al. [29] on predictive modeling, underscores our strong background in applying LLMs to various real-world contexts. Additionally, our research on fairness, including the forthcoming paper[5] "FairUS: UpSampling Optimized Method for Boosting Fairness," highlights our ability to address issues of bias and fairness in machine learning. Our expertise in handling language-specific challenges, such as the morphological complexity of Hebrew [9, 30], further equips us to effectively mitigate bias in diverse LLMs. We are also well-versed in utilizing new reinforcement learning frameworks [40], expanding our capacity to innovate in this area. Our

[5] `https://www.ecai2024.eu/programme/accepted-papers`

familiarity with Hebrew corpora, Hebrew-specific LLMs, and linguistic structures will be crucial in developing culturally sensitive methods for identifying and reducing biases.

Cloud-based computational resources will be used to manage the significant processing demands involved in analyzing LLMs. Access to major LLM APIs and open-source models allows us to experiment with a wide range of systems, while large-scale cloud platforms support the training and evaluation of models at scale. These resources, combined with datasets we have collected from LLM prompts and responses, ensure that our research will benefit from diverse linguistic and cultural contexts.

7 Expected results and potential pitfalls

The rapid emergence of new LLMs could introduce variability and require regular re-evaluation of our bias detection and mitigation strategies. We note that we have already demonstrated the generalizability of our method across multiple LLMs, ensuring its continued effectiveness as new models are introduced. However, the ongoing evolution of architectures and data sources means that we will need to regularly test and refine our methods to prevent the introduction of unforeseen biases. Another critical factor is the need for labeled data, which can be resource-intensive to gather. We have developed a process for collecting labeled responses with high inter-annotator agreement, which ensures the reliability and accuracy of our labels. Despite this, scaling the process to handle larger datasets and account for diverse demographic groups may introduce challenges. Automating aspects of the labeling process, while maintaining quality control, could offer a potential solution in the future.

The expected results of this research are multifaceted. First, we aim to reduce disparities in LLM responses across various demographic groups, such as gender, age, and ethnicity, leading to more equitable outputs that ensure fair treatment for users from diverse backgrounds. We also expect to demonstrate that our bias mitigation strategies are adaptable across different LLM architectures and languages, including models trained in rich morphological languages like Hebrew.

This research will lay the foundation for a formalized framework for bias mitigation in LLMs. The suggested strategies can be adopted by other researchers, contributing to ongoing efforts in AI ethics and fairness. In doing so, the research will provide practical, adaptable, and ethically sound solutions to ensure fairness across a wide range of models, languages, and demographic contexts.

References

[1] AI@Meta. Meta - llama 3, 2024.

[2] Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? *arXiv preprint arXiv:2406.10486*, 2024.

[3] Anthropic. Mar 2024.

[4] Liat Ayalon and Clemens Tesch-Römer. Introduction to the section: Ageism—concept and origins. *Contemporary perspectives on ageism*, pages 1–10, 2018.

[5] Yanhong Bai, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xingjiao Wu, and Liang He. FairMonitor: A Dual-framework for Detecting Stereotypes and Biases in Large Language Models. *arXiv e-prints*, page arXiv:2405.03098, May 2024.

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

[7] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[8] Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1):34–48, 2024.

[9] Seffi Cohen, Edo Lior, Moshe Bocher, and Lior Rokach. Improving severity classification of hebrew pet-ct pathology reports using test-time augmentation. *Journal of Biomedical Informatics*, 149:104577, 2024.

[10] Seffi Cohen, Dan Presil, Or Katz, Ofir Arbili, Shvat Messica, and Lior Rokach. Enhancing social network hate detection using back translation and gpt-3 augmentations during training and test-time. *information Fusion*, 99:101887, 2023.

[11] Shir Etgar, Gal Oestreicher-Singer, and Inbal Yahav. Implicit bias in llms: Bias in financial advice based on implied gender. *Available at SSRN*, 2024.

[12] Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *arXiv preprint arXiv:2306.15087*, 2023.

[13] Matthew Freestone and Shubhra Kanti Karmaker Santu. Word embeddings revisited: Do llms offer something new? *arXiv preprint arXiv:2402.11094*, 2024.

[14] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[15] Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. *arXiv preprint arXiv:2203.12574*, 2022.

[16] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv e-prints*, page arXiv:2310.06825, October 2023.

[17] Chayanan Kerdpitak and Kittisak Jermsittiparsert. Impact of gender-based, age-based, and race-based discrimination on satisfaction and performance of employees. *Systematic Reviews in Pharmacy*, 11(2), 2020.

[18] Hadas Kotek, Rikker Dockum, and David Q Sun. Gender bias and stereotypes in large language models. *arXiv preprint arXiv:2308.14921*, 2023.

[19] Mosh Levy, Shauli Ravfogel, and Yoav Goldberg. Guiding LLM to fool itself: Automatically manipulating machine reading comprehension shortcut triggers. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8495–8505, Singapore, December 2023. Association for Computational Linguistics.

[20] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*, 2023.

[21] Gallil Maimon and Lior Rokach. A universal adversarial policy for text classifiers. *Neural Networks*, 153:282–291, 2022.

[22] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[23] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1953–1967, Online, November 2020. Association for Computational Linguistics.

[24] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2), 6 2023.

[25] Peter Bjorn Nemenyi. *Distribution-free multiple comparisons.* Princeton University, 1963.

[26] Valentina Pyatkin, Frances Yung, Merel CJ Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. Design choices for crowdsourcing implicit discourse relations: revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11:1014–1032, 2023.

[27] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

[28] Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. Mrl parsing without tears: The case of hebrew. *arXiv preprint arXiv:2403.06970*, 2024.

[29] Guy Shtar, Asnat Greenstein-Messica, Eyal Mazuz, Lior Rokach, and Bracha Shapira. Predicting drug characteristics using biomedical text embedding. *BMC bioinformatics*, 23(1):526, 2022.

[30] Adir Solomon, Amit Magen, Simo Hanouna, Mor Kertis, Bracha Shapira, and Lior Rokach. Crime linkage based on textual hebrew police reports utilizing behavioral patterns. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2749–2756, 2020.

[31] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[32] Naznin Tabassum and Bhabani Shankar Nayak. Gender stereotypes and their impact on women's career progressions from a managerial perspective. *IIM Kozhikode Society & Management Review*, 10(2):192–208, 2021.

[33] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*, 2024.

[34] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, and Jiahui Yu. Gemini: A family of highly capable multimodal models, 2023.

[35] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, and Laurent Sifre. Gemma: Open models based on gemini research and technology, 2024.

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[37] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*, 2023.

[38] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.

[39] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. "as an ai language model, i cannot": Investigating llm denials of user requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2024.

[40] Chen Yanai, Adir Solomon, Gilad Katz, Bracha Shapira, and Lior Rokach. Q-ball: Modeling basketball games using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8806–8813, 2022.

[41] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

[42] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

[43] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*, 2023.