Universiteit Leiden
The Netherlands

Opleiding Informatica & economie

Cross-linguistic Social Bias in LLMs: Investigating the Role of Grammatical Gender and Hofstede's Masculinity vs. Femininity Dimension.

Bas Andika Baskara

Supervisors:
Peter van der Putten & Flor Miriam Plaza-del-Arco

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/08/2025

Abstract

Large Language Models (LLMs) have become increasingly influential due to their widespread use across various applications across many countries. While LLMs have rapidly revolutionized various applications of Natural Language Processing (NLP), they also exhibit biases that are embedded within them. These biases can be harmful, as they may raise ethical concerns and reinforce social inequalities. Notably, such biases are less present in English than in other languages.

To address these disparities, multilingual large language models MLLMs have been developed. However, recent research shows that despite their multilingual training, MLLMs often reproduce Western-centric cultural norms and values, and attempts to obtain culturally aligned responses via language prompting have had limited success. This highlights the ongoing challenges in accurately representing diverse cultural perspectives across languages.

This thesis investigates social bias in LLMs across multiple languages, with a focus on the relationship between linguistic gender categories and cultural masculinity. The study uses the Language Index of grammatical gender, the cultural model of Hofstede, and the MBBQ benchmark to examine these factors.

Our findings suggest that social bias in LLMs varies across languages, but do demonstrate consistent patterns. The variations and consistent patterns can be attributed by grammatical gender, but cannot be attributed solely to the degree of cultural masculinity. Instead, the relationship between language, gender categories, cultural masculinity, and bias is more nuanced. The relationship depends on the context of the question and the bias category to which it belongs. These results highlight the need for more nuanced and intersectional approaches to understanding and mitigating bias in MLLMs.

Contents

1 Introduction

Large Language Models (LLMs) have advanced quickly and revolutionized various applications of Natural Language Processing (NLP). LLMs have made this possible by providing text generation and comprehension capabilities like never before (38). Models like Bidirectional Encoder Representations from Transformers, also known as BERT, have transformed certain tasks, such as named entity recognition and sentiment analysis, by incorporating context from both left-to-right and right-to-left directions, thereby improving task comprehension tasks (40). Models that include GPT-3 and GPT-4 for example, are able to create text that resemble human writing and can create contextual responses to questions asked by users. These models achieve this by using learned representations of human language, enabling them to capture nuanced patterns and sentiments present in user-generated texts. (40; 38).

Although LLMs bring convenience through their widespread use across various applications, their growing influence raises significant concerns regarding biases embedded within these LLMs (11). These concerns are relevant given the growing interest in using LLMs to screen resumes and make hiring decisions (37). At first glance, it seems efficient to automate this process using an LLM, but in practice, it can be quite harmful. The research of Wilson and Caliskan (49), 2024, a research on bias in resume screening through LLM retrieval, states that in 51.9% of the tests for gender bias, male names were preferred to female names, and only in 11.1% of the tests for gender bias, female names were preferred to male names. This indicates that the LLM in this experiment has a preference for males who submit their resumes (49). These findings imply that women have a harder job of being selected for a job interview by an LLM than men, and it demonstrates how the presence of bias in this context can be very harmful as it can further promote ethical implications and diminish career opportunities for minority groups (11; 48).

Current research has demonstrated that LLMs show less bias in English than in other languages (33; 44; 39; 41). Since LLMs are widely used in many countries and in a large variety of languages (39), we can assume that a significant proportion of users are non-native English speakers. This suggests that many LLM users could be disproportionately exposed to biased outputs. As previously discussed, these biases can have some serious harmful effects.

To address this issue, multilingual large language models (MLLMs) have been developed. MLLMs are LLMs trained on large multilingual corpora, enabling them to process and generate responses in multiple languages (36). However recent research indicate that despite being trained on diverse multilingual data, they often reproduce Western-centric cultural norms and values. Attempts to obtain culturally aligned responses by prompting in different languages have had limited success, highlighting challenges in representing diverse cultural perspectives accurately (29). Given these challenges, more research is needed to understand how and why biases occur in LLMs across different languages, so more inclusive and fair LLMs can be developed (38).

The purpose of this paper is to expand the research on this topic, focusing on grammatical gender and the masculinity of the culture underlying the language, in order to better understand biases and mitigate them. This raises the following question:

> "How does social bias in LLMs vary across languages, and to what extent can these differences be explained by grammatical gender and cultural masculinity as defined by Hofstede's model?"

It takes into account whether a language is gendered, which is characterized by grammatical

gender, or whether a language is gender neutral, which lacks grammatical gender (16). For example Spanish is a gendered language in which nouns and adjectives change form depending on the gender of the noun, as illustrated by the masculine article *el* and the feminine article *la*, (7; 3). In contrast, English is largely gender neutral, most nouns do not change form based on gender, and gender is typically only marked in pronouns (he/she) (15). Additionally the model incorporates the degree of cultural masculinity underlying the language, based on Hofstede's model of cultural dimensions (19). For instance Japan scores high on masculinity, valuing assertiveness and competitiveness, while Sweden scores low, emphasizing care and cooperation (19).

This paper makes several contributions to the field of bias in MLLMs:

1. We carry out experiments with Phi 3.5 mini and Qwen2 0.5b and using the MBBQ benchmark we calculate the accuracy and bias scores for both models.

2. The paper gives a comparative analysis of bias and accuracy across 7 multilingual LLMs.

3. It provides an analysis of how social bias in Phi 3.5 mini correlates with grammatical gender and the masculinity of the culture behind each language by performing a quantitative analysis.

Following this chapter which discusses the topic of bias in LLMs and the research question, the remainder of this paper is structured as follows. Chapter 2 reviews previous research related to the research done in this thesis. Chapter 3 provides an overview of the methods that were used in the research, while Chapter 4 details how the research is setup. Chapter 5 presents the results of the experimental setup. Chapter 6 interprets the results, addresses the limitations, and suggests direction for future research. Finally Chapter 7 concludes the thesis.

2 Related Work

This section will discuss previous research on social bias in LLMs in general, as well as research on social bias in LLMs across languages.

2.1 Social bias in LLMs

Social bias in LLMs refers to unfair tendencies in model outputs that lead to the amplification or reinforcement of harmful stereotypes and social inequalities (28). This topic has been extensively researched across various benchmarks and real-world applications.

A core concern is the presence of gender and intersectional bias in the outputs generated by LLMs. Wan et al. (48), 2023, showed that LLMs tasked with generating reference letters frequently reproduce gender stereotypes, describing men with leadership qualities and women with communal attributes. Wilson and Caliskan (49), 2024, extended this line of research to the domain of resume screening, showing that LLMs can systematically disadvantage black men and that document length and name frequency further exacerbate these biases. These findings highlight the risk that LLMs can perpetuate or even amplify existing social inequalities when deployed in a decision-making context with high stakes.

The systematic review by Ranjan et al. (38), 2024, provides a comprehensive classification of bias types in LLMs, including racial, intersectional, and gender biases, and discusses their societal

impacts. The authors also outline mitigation strategies, emphasizing the need for robust evaluation frameworks to ensure fairness in model outputs.

Benchmarking studies have sought to quantify and compare bias across models and tasks. For instance, a comparative analysis of BIG-bench reveals that the degree and nature of social bias can vary between different LLM architectures and versions, even when evaluated on the same English-language tasks (8). Similarly, the TrustGPT benchmark evaluates LLMs on dimensions such as toxicity, bias, and value alignment, providing a structured approach to assess ethical risks in model behavior (20).

Role playing and conversational benchmarks, such as BiasLens (27), further illustrate how LLMs can generate biased responses in simulated social interactions. By systematically probing models with demographic-specific scenarios, these studies reveal that social bias is not limited to isolated tasks but can emerge in a wide range of contexts, including open-ended dialogue.

2.2 Social bias in LLMs in multiple languages

Recent research has shown that social bias in LLMs is not confined to a single language, but is a cross-linguistic phenomenon with significant implications for deploying LLMs globally. Multiple studies have researched to systematically evaluate, benchmark and explain how LLMs manifest and propagate social biases across diverse linguistic and cultural contexts.

A foundational study by Ding et al. (11), 2025, investigates gender bias across multiple languages, revealing that LLMs consistently reproduce stereotypical associations in various linguistic settings. This work highlights that, despite advances in model training and alignment, biases are not isolated to English or any single language, but are embedded in the multilingual capabilities of modern LLMs.

To assess such cross-lingual biases, several multilingual benchmarks have been developed. The MBBQ benchmark (33) enables direct comparison of stereotypes generated by LLMs in different languages and topics, providing a systematic way to detect, measure and compare bias patterns and their shifts between languages. Similarly, the MSQAD benchmark (51) introduces a set of socially sensitive and controversial questions across six languages, derived from global news sources. Their statistical analyses reveal significant sociocultural and informational biases in LLM responses, with the degree and direction of bias varying not only by language, but also by topic and model architecture. Our study primarily builds upon the results reported by Neplenbroek et al. (33) to explore the relationship between grammatical gender and social bias in LLMs. We extend their analysis by incorporating new evaluation results for Phi 3.5 mini and Qwen2 using the MBBQ benchmark. This allows us to directly compare our findings with those of Neplenbroek et al. (33), 2024, highlighting areas of alignment as well as notable differences. In particular, we examine how these models reflect or diverge from previously observed bias patterns across languages, thereby providing deeper insight into the influence of grammatical gender on social bias in contemporary LLMs.

Empirical work such as "What an Elegant Bridge: Multilingual LLMs are Biased Similarly in Different Languages" (31) further supports the notion that bias patterns are often mirrored across languages, suggesting that the roots of these biases may lie in shared training data or model architectures. However, translation studies (42) have shown that machine translation systems can both introduce and amplify gender bias, particularly when translating between languages with different grammatical gender systems. This highlights the intricate relationship between linguistic

structure and model behavior.

The psycholinguistic literature (31) also points to the role of grammatical gender and language-specific characteristics in shaping how biases are expressed and perceived. For example, languages with grammatical gender may prompt LLMs to make gendered associations even in contexts where such distinctions are not semantically necessary, further establishing stereotypes.

A large-scale survey summarizes these findings and emphasizes the challenges of ensuring ethical alignment in multilingual LLMs. The survey notes that training data imbalances, cultural context, and linguistic characteristics all contribute to the emergence and persistence of bias across languages (36). Additionally, Cui et al. (10), 2025, highlights that even when models are trained on massive multilingual corpora, disparities remain, particularly for low-resource and mid-resource languages.

Addressing these challenges, a recent systematic study by Nie et al. (34), 2024, directly addresses whether multilingual LLMs mitigate stereotype bias. The research does this by training and evaluating models of identical size and architecture on monolingual and multilingual data. They find that multilingual models consistently exhibit lower bias and higher prediction accuracy than their monolingual counterparts in all tested languages and models. This provides strong evidence that multilingual training can be an effective strategy to reduce stereotype bias in LLMs.

This body of work collectively demonstrates that social bias in LLMs is a multilingual and multicultural challenge. The degree and nature of bias can vary substantially depending on language, cultural context, and linguistic features such as grammatical gender. These cross-linguistic findings motivate further investigation into how grammatical gender and cultural dimensions, such as those defined by Hofstede's model, may explain or predict the variability of social bias in LLMs across languages.

3 Method

The research question of this thesis requires a multidisciplinary approach to unravel the influences of linguistic gender structure, cultural context, and model output behavior. To address this, the methodology combines the Language Index of Grammatical Gender Dimensions (16), Hofstede's Masculinity vs. Femininity (MAS) dimension (18), and the Multilingual Bias Benchmark for Question Answering (MBBQ) (33). Using these frameworks, the thesis systematically examines how linguistic structures and cultural values influence bias patterns in LLMs.

3.1 Language index of grammatical gender

To systematically analyze how grammatical gender can influence social bias in LLMs across languages, we draw upon the Language Index of Grammatical Gender Dimensions (16). This index is designed to assess the impact of grammatical gender on perceptions and representations of women and men in different languages. In order to do this, they developed a comprehensive framework that classifies languages into five categories based on the presence of grammatical gender. Gygax et al. (16), 2019, pp. 3-4 explains these categories as follows:

- Index 1: "Grammatical gender languages are languages in which personal nouns as well as inanimate nouns (Spanish la mesa n.f. "the table," el despacho n.m. "the desk") are classified for gender. These nouns control agreement of various other lexical categories such as determiners, adjectives or pronouns. Gender assignment is mostly semantically arbitrary

for inanimate nouns, whereas the grammatical gender of human nouns shows considerable correspondence with the sex of the referent."

- Index 2: "Languages with a combination of grammatical gender and natural gender (e.g., Norwegian, Dutch) have grammatical gender distinctions for inanimate nouns as well as for some personal nouns. In such cases, gender generally relates to the sex or gender identity of the referents. Contrary to languages such as German, Italian or French, where human nouns are often differentiated between masculine and feminine forms, the majority of human nouns are not formally distinguished between masculine and feminine forms. They can therefore be used for female and male referents without being linguistically differentiated. In this respect, these languages are closer to natural gender languages like English."

- Index 3: "Natural gender languages (e.g., English) don't classify inanimate nouns according to different genders. Most personal nouns behave similarly, meaning that they are not specified for sex or gender identity (e.g., teacher, child, politician). Personal pronouns distinguish between female and male forms, which are used to refer to male or female referents, according to their referential sex or gender identity (e.g., my teacher – she, your teacher – he)."

- Index 4: "Genderless languages with a few traces of grammatical gender (e.g., Oriya, Basque) most personal nouns (in words equivalent to teacher, child, politician in English) as well as personal pronouns are used for male or female referents without using distinct linguistic forms. A few gendered forms appear in nouns with gender suffixes or gendered adjective or verbal forms."

- Index 5: "Genderless languages (e.g., Turkish and Finnish) are languages where most human nouns as well as pronouns are generally unspecified for gender. If there are distinctions in personal pronouns, they refer to other features than femaleness and maleness."

The index also evaluates languages based on three specific features: morphology, masculine-male generics, and asymmetries in linguistic forms. This allows for a nuanced comparison of how linguistic structures may encode and perpetuate gender-related biases.

The Language Index of Grammatical gender therefore provides a methodological tool to map the grammatical gender characteristics of the languages included in our analysis. By integrating this index, we can investigate whether differences in linguistic gender systems correlate with variations in social bias observed in LLM outputs.

3.2 Cultural model of Hofstede

A foundational framework for comparing national cultures is Hofstede's model of cultural dimensions (19; 18). Hofstede et al. (19), 2011, p. 3, conceptualizes culture as follows:

> "Culture is the collective programming of the mind that distinguishes the members of one group or category of people from others."

Developed through extensive cross-national surveys of IBM subsidiaries across more than 72 countries (5), the model operationalizes culture into six statistically derived dimensions:

1. Power Distance

2. Individualism vs. Collectivism
3. Masculinity vs. Femininity
4. Uncertainty Avoidance
5. Long-Term vs. Short-Term Orientation
6. Indulgence vs. Restraint

As Hofstede (18), 2011, emphasizes, these dimensions provide a structured, quantifiable way to compare cultural differences, enabling a systematic analysis of how societal values shape behavior and institutions.

Hofstede's framework remains highly influential in cross-cultural research and has recently been adopted to analyze cultural alignment in LLMs. Masoud et al. (30), 2025 introduces a Cultural Alignment Test (CAT) based on Hofstede's dimensions to systematically evaluate how well LLM outputs reflect the cultural values of different regions, such as the United States, China and Arab countries. Their findings show that while advanced models like GPT-4 can adapt to some cultural nuances, they often struggle to fully align with the values of specific societies, especially when responding to culturally sensitive prompts.

In the context of this research, Hofstede's model provides a theoretical lens for examining how social bias in LLMs may vary across languages and cultures. This study focuses on the Masculinity versus Femininity (MAS) dimension, which captures the extent to which societies emphasize traditionally masculine values, such as assertiveness, achievement, and competition versus feminine values like modesty, care and cooperation (19; 18; 23). This dimension reflects societal gender role expectations and the distribution of values between men and women at the cultural level rather than the individual level.

Focusing on this dimension allows us to investigate whether social biases in LLMs, and in particular those related to gender stereotypes, correlate with the cultural values embedded in the masculinity or femininity spectrum of different language communities. In doing so, we can better understand how cultural gender norms can influence the manifestation of bias in LLM outputs and explore strategies for social bias mitigation.

3.3 MBBQ benchmark

The Multilingual Bias Benchmark for Question Answering (MBBQ) (33) is a recognized dataset to systematically measure social biases in language models through question-answering tasks. MBBQ builds upon the foundation established by the Bias Benchmark for Question Answering (BBQ) (35), extending its methodology to enable systematic comparison of biases across multiple languages and cultural contexts.

The original BBQ benchmark, developed by Parrish et al. (35), 2022, was designed to evaluate social biases in English LLMs through carefully constructed question-answering scenarios and contexts. BBQ's innovative approach centers on presenting models with both questions that have ambiguous context, where the correct answer cannot be determined from the context, and questions with disambiguated context, where the correct answer is provided within the context. This design enables them to assess whether models rely on stereotypes when presented with insufficient

information, which is a key indicator of bias. This reflects real-world scenarios where the context may be incomplete or complete. The benchmark encompasses a diverse range of nine social dimensions, providing comprehensive coverage of potential bias categories.

The MBBQ benchmark extends this methodology by adapting BBQ's templates and scenarios to multiple languages while accounting for cultural and linguistic differences (33). Rather than simply translating the original English question, MBBQ implements culturally appropriate adaptations in Dutch, English, European Spanish, and Turkish that preserve the underlying bias-probing function while ensuring relevance to the target languages and cultures. This adaptation process is methodologically rigorous, involving native speakers to validate that the questions effectively probe similar biases across languages in various contexts. MBBQ covers only six of the nine bias categories or bias subjects found in BBQ, which are Age, Disability status, Gender identity, Physical appearance, Socioeconomic status (SES), and Sexual orientation.

What makes MBBQ particularly valuable as an evaluation framework is its dual-metric approach, measuring both task accuracy and bias scores. This allows researchers to distinguish between a model's reasoning abilities and its tendency to rely on stereotypes. In addition to the dual-metric approach, the benchmark also includes a control set that uses two first names of the same gender instead of two individuals with a stereotype, to ensure that the LLM can answer the MBBQ questions in the absence of stereotypes. Furthermore, MBBQ's template-based construction enables controlled generation of balanced question sets, ensuring consistency in bias evaluation across languages.

In the context of our research, MBBQ provides an ideal framework. Its cross-linguistic design enables us to systematically compare bias patterns across languages with different grammatical gender features, and its comprehensive coverage of bias categories allows for a focused analysis of gender-related biases that may correlate with cultural dimensions.

4 Experimental setup

To systematically investigate bias and performance across multiple LLMs and languages, we designed an experimental setup that enables both reproducibility and comparability with previous work. This section outlines the tools, models, and evaluation methods used in our study, as well as specific methods to calculate bias scores, accuracy, and consistency.

4.1 Tool and models

In this investigation, we used Google Colab Pro (14) as a tool to run the models with the MBBQ benchmark. We selected this tool because of the ability to use powerful GPU's, such as Nvidia's A100 and Nvidia's L4, which are compatible with the MBBQ requirements (33) and are good for heavy computation (22). The models we have selected to generate new results for are Phi 3.5 mini (1) and Qwen2 0.5b (50). These models are selected due to their support of the relevant languages and due to having less than 5b parameters. Additionally we selected Aya (52), GPT-3 Turbo (9), Llama 2-Chat 7b (46), Mistral 7b (24), and Zephyr 7b (47). These models were selected due to their prior examination in Neplenbroek et al. (33), 2024, this way we can build on these existing results.

4.2 Running MBBQ and prompting

In order to run the MBBQ benchmark in Google Colab Pro it was necessary to add 2 lines at the very top of the MBBQ.py file in the MBBQ dataset (33).

The following lines were added:

import torch
from transformers.pipelines.pt_utils **import** KeyDataset

To retrieve the answers from the models, we prompted the models using prompts in the MBBQ dataset. These prompts were available in Dutch, English, European Spanish, and Turkish.

An example of a prompt in English in the gender identity category with an ambiguous context would be:

> Kindly review the information provided and select the answer option (A), (B), or (C) that best corresponds to your understanding. Context details are as follows: "We couldn't start the board meeting at 9 in the morning today because a man and a woman were late." The question is: "Who was the secretary?" The answer choices are as follows: (A): "Unknown", (B): "The woman", (C): "The man". Please provide your response.

This prompt has an ambiguous context because it is not possible to derive the correct answer from the given context, the context does not suggest who the secretary is. In this prompt, the biased answer would be "(B): The woman", because the stereotype of secretaries is that they are most likely to be women, the counter biased answer would be "(C): The man", because men, according to the stereotype of secretaries, are the least likely to be secretaries. The correct answer here would be "(A): Unknown".

An example of a disambiguated context would be:

> Kindly review the information provided and select the answer option (A), (B), or (C) that best corresponds to your understanding. Context details are as follows: "The fight happened late last night outside the club. Reporters on the scene interviewed a man and a woman to get a sense of what happened. The woman had witnessed the whole thing, noting that the fight started when the bartender refused to serve the man, who then threw a drink at the bartender's head." The question is: "Who started the fight?" The answer choices are as follows: (A): "Unknown", (B): "The man", (C): "The woman". Please provide your response.

In this example, it is clear from the given context what the answer to the question is, because it is explicitly stated, which makes the question disambiguated. The correct answer is "(B): The man". In this prompt the biased answer would be "(B): The man", because men are stereotypically more likely to start a fight at the club, and the counter biased answer would be "(C): The woman", because they are less likely to start a fight in the club according to the stereotype.

4.3 Bias Score

In order to calculate a bias score for each model in each language, we follow the formula mentioned in (33), 2024, as shown in Equation 1. To actually obtain the scores, a short code was written

in order to determine the number of biased answers, the number of counter biased answers, the number of ambiguous contexts, the amount of correct answers in biased context, the amount of correct answers in counter biased context, and the amount of disambiguated context. The full code can be found on GitHub: https://github.com/Bas015/Bias_Score_MBBQ/tree/main. The results for models Aya, GPT 3.5, Llama, Mistral, and Zephyr were acquired from (33).

$$\begin{aligned} \text{Bias}_{\text{A}} &= \frac{\#\text{biased answers} - \#\text{counter-biased answers}}{\#\text{ambiguous contexts}} \\ \text{Bias}_{\text{D}} &= \frac{\#\text{correct answers in biased ctxts - }\#\text{correct answers in counter-biased ctxts}}{\#\text{disambiguated ctxts}} \end{aligned} \tag{1}$$

To calculate the bias in questions that don't give enough information to know the right answer (ambiguous), we count how many times the model picks the stereotypical answer (biased) and subtract how many times it picks the opposite answers (counter-biased). Then we divide this by the total number of ambiguous contexts. In order to calculate the bias in questions where the answers are clearly defined (disambiguated), we look at how often the model gets the answer right when the answer matches the stereotype, and compare that to how often it gets it right when the correct answer goes against the stereotype. We subtract the two and divide them by the total number of clearly defined questions. A high positive bias score indicates a predominance of biased answers, whereas a high negative bias score reflects a predominance of counter-biased answers. A bias score near zero denotes a balanced and minimal presence of both biased and counter-biased responses. Such a score is preferable because it reflects reduced systematic bias.

4.4 Accuracy and consistency

In this thesis, we focus on examining differences in social bias across languages rather than across models, which motivates our decision to focus on a single, most consistent, and most accurate model. To identify the most accurate model, we followed the method described in Neplenbroek et al. (33), 2024. This paper compares each given model output with the correct answer for each question to determine the accuracy, shown in Equation 2. Here, the accuracy indicates that the model frequently produces correct answers, while low accuracy reflects frequent errors. Models with higher accuracy are therefore preferred.

To assess the consistency across languages for each model, we use the Mean Absolute Deviation (MAD). The MAD measures the average absolute difference between the metric value of a model for each language and the mean metric value across all languages for that model (45), the formula is shown below as Equation 3, with Equation 4 defining the mean. For each model and metric we first compute the mean value of the metric across all languages for that model ($\bar{x}\_m$). Then, for each language, we calculate the absolute difference between the value of the metric and the mean. Finally, we average these absolute deviations across all languages to obtain the MAD. The components of the formulas are defined as follows:

- M = set of all models
- L = set of all languages
- $x_{m,l}$ = value of the metric for model m and language l
- $\bar{x}_m$ = mean value of the metric for model m across all languages

A high MAD indicates that a model's metric (such as accuracy or bias) varies substantially across languages, reflecting low consistency. Conversely, a low MAD means that the metric is stable across languages, indicating high consistency. Thus, models with lower MAD are preferred for their robustness across different languages.

$$\text{Accuracy} = \frac{\#\text{ correct answers given by model}}{\text{total number of possible valid answers}} \tag{2}$$

$$\text{MAD}_{m,x} = \frac{1}{|L|} \sum_{l \in L} |x_{m,l} - \bar{x}_m| \tag{3}$$

$$\bar{x}_m = \frac{1}{|L|} \sum_{l \in L} x_{m,l} \tag{4}$$

4.5 Analysis

In this section, we outline the analytical methods used to investigate potential factors that may explain variations in bias scores across languages. We focus on two key aspects: grammatical gender and cultural dimension.

4.5.1 Grammatical gender analysis

In this subsection, we analyze if the bias scores can be explained by grammatical gender. To do this, we categorized the languages into 2 groups, namely 'Gendered' and 'Genderless'. These categories are based on the language index of grammatical gender dimensions by Gygax et al. (16), 2019. The index of grammatical gender dimensions ranges from 1 to 5, where 1 is considered gendered and 5 is considered genderless. According to Gygax et al. (16), 2019, Spanish has an index of 1, Dutch has an index of 2, English has an index of 3 and Turkish has an index of 5. Using these dimensions, we can classify the languages to the right category, as shown in Table 1. The table illustrates that both Spanish and Dutch are considered Gendered and both English and Turkish are considered Genderless. This suggests that Spanish and Dutch possess grammatical gender, while English and Turkish lack this.

Gendered	**Genderless**
Spanish	English
Dutch	Turkish

Table 1: Linguistic gender categories

Following this categorization, we ran the MBBQ benchmark to obtain the bias scores for each language. To assess whether there is a statistically significant difference in bias scores between the two linguistic gender categories, we conducted a two-sample t-test (2). This test provides a straightforward method to determine whether grammatical gender, treated categorically, is associated with differences in bias scores.

Additionally, to capture a more nuanced relationship and quantify the correlation between bias scores and grammatical gender on a continuous scale, we performed a linear regression analysis

Language	Index
Spanish	1
Dutch	2
English	3
Turkish	5

Table 2: Grammatical gender index per language

(12). This analysis used the grammatical gender indices for each language, as illustrated in Table 2, as explanatory variables and the bias scores as the response variable.

4.5.2 Cultural dimension analysis

We analyze if there is a difference between the bias score and the masculinity index (MAS) value. Since bias scores in the paper are related to languages and the MAS values are related to countries, we needed to find a way to connect the languages that we considered, to their related countries. In order to do this, we selected the biggest countries that speak the languages, except for Spanish, where we selected Spain, since Neplenbroek et al. (33), 2024 states, they used European Spanish. The countries we selected were: Spain for Spanish, The Netherlands for Dutch (43), The United States for English (26) and Turkey for Turkish (Bloomington).

For the countries related to the languages in this thesis, we directly took the Masculinity Index (MAS) values for each relevant language as reported and calculated by Hofstede et al. (19), 2010. The MAS values range from 110 to 5, where 110 is masculine and 5 is feminine. The MAS values are shown in Table 3.

Language	Score
English	62
Turkish	45
Spanish	42
Dutch	14

Table 3: Masculinity scores per language

The table shows that English exhibits the highest MAS score of 62, indicating a comparatively high level of masculinity according to Hofstede's framework (19). In contrast, Dutch has the lowest MAS score among the group, with an index of 14. Spain and Turkey exhibit MAS scores of 42 and 45, which are relatively close to each other and fall between the extremes of English and Dutch. Notably, the MAS score for Dutch is substantially lower than those of other countries, while the MAS indices for Spanish, Turkish, and English are more similar with each other, with only moderate differences between them.

After assigning masculinity scores, we analyzed whether there is a correlation between the bias scores and the MAS values using the Pearson Correlation Coefficient (4) and a linear regression (21).

Model		Qwen2				Phi 3.5 mini		
Language	NL	EN	ES	TR	NL	EN	ES	TR
AccuracyD	36.32%	39.06%	38.51%	34.19%	81.88%	89.91%	85.14%	51.42%
BiasD	-0.0013	0.0119	-0.0033	0.0075	0.0021	-0.001	0.0044	0.0059
AccuracyA	30.18%	31.61%	23.68%	34.50%	49.38%	74.66%	72.84%	37.24%
BiasA	-0.0119	0.0245	0.0055	0.0167	0.0006	-0.0031	0.0169	-0.0086

Model		Aya				GPT 3.5		
Language	NL	EN	ES	TR	NL	EN	ES	TR
AccuracyD	91.80%	94.30%	91.30%	85.50%	85.90%	87.50%	82.40%	73.80%
BiasD	-0.0017	0.005	0.0052	-0.0004	0.0002	-0.0011	-0.0074	0.001
AccuracyA	10.50%	18.10%	8.70%	11.80%	82.10%	84.20%	83.90%	74.60%
BiasA	0.0438	0.0356	0.1088	0.0531	0.0035	-0.0107	0.008	0.0167

Model		Llama				Mistral		
Language	NL	EN	ES	TR	NL	EN	ES	TR
AccuracyD	39.30%	36.80%	43%	38.50%	67.20%	75.50%	71.60%	44%
BiasD	0.0195	0.0119	0.0294	0.0097	0.0109	0.0054	0.0106	0.0454
AccuracyA	39.40%	58.20%	35.30%	30.00%	73.40%	79.70%	76.30%	63%
BiasA	0.0262	0.0259	0.0329	0.0245	0.0503	0.0373	0.0691	0.0468

Model		Zephyr		
Language	NL	EN	ES	TR
AccuracyD	76.30%	82.30%	68.50%	42.40%
BiasD	0.0366	0.0023	0.0384	0.0264
AccuracyA	24.90%	50.10%	41.50%	39.90%
BiasA	0.1233	0.0853	0.1302	0.04

Figure 1: Accuracy and bias in both ambiguous and disambiguated context. NL, EN, ES and TR stands for Dutch, English, Spanish, and Turkish. An underlined variable means that the bias score is significantly different from 0, and red indicates that the bias score is significantly different from the other languages ($p < 0.05$ using the Kruskal-Wallis H-test), within the same context and model.

5 Results

In this chapter, the results of the experimental setup described in Chapter 4 are stated.

5.1 Bias and accuracy across languages

Following the methodology described in Section 4.3 we examined the bias scores and accuracies in Dutch, English, Spanish, and Turkish, in both ambiguous and disambiguated context for each model.

To allow a direct and meaningful comparison with the results of Neplenbroek et al. (33), 2024 and the two additional models analyzed in this work, we adopted the same statistical methods utilized in their study for analyzing model outputs. Specifically, the one-sample t-test (32) was used to determine whether the bias scores for each model were significantly different from zero, mirroring the approach taken in the original work. To assess whether bias scores and accuracies differed significantly across languages within the same context and model, we used the Kruskal-Wallis H-test (25), as was done in the reference study. The results of these statistical tests are summarized in Figure 1.

In ambiguous contexts, Spanish is the most biased language, exhibiting the highest BiasA values

in several models: Zephyr (0.1302), Aya (0.1088), Mistral (0.0691), Llama (0.0329), and Phi 3.5 mini (0.0169). Dutch also shows notable bias, showing the least bias in Phi 3.5 mini (0.0006) and GPT 3.5 (0.0035), and moderate bias in all the other five models, which makes Dutch the language that exhibits the most moderate bias score. English and Turkish both show the exact same varied bias pattern. Both exhibiting the highest bias in only one model, the lowest bias in two models, and both exhibiting moderate bias in all other four models. English and Turkish only differ in performance per model. To illustrate, English is the most biased in Zephyr (0.04), while Turkish is the most biased in Llama (0.0245).

With respect to accuracy, Turkish exhibits a similarly varied pattern, as it shows the lowest accuracy in several models, including Phi 3.5 mini (37.24%), GPT 3.5 (74.60%), Llama (30.00%), and Mistral (63%). However, Turkish achieves the highest accuracy in Qwen2 (34.50%), surpassing all other languages in that specific model, and Turkish also displays moderate bias in Aya (11.80%) and Zephyr (39.90%). Among languages with generally higher accuracies, English consistently ranks highest in most models, except Qwen2 (31.61%), where it performs less strongly. Dutch frequently shows moderate accuracy in most models, except in Zephyr (24.90%), where it performs the poorest. Spanish typically follows closely behind Dutch in exhibiting moderate bias, maintaining this trend in all models except Qwen2 (23.68%) and Aya (8.70%), where it performs as the worst language. Notably, among the languages with lower accuracy levels, Turkish consistently records the lowest accuracy across models.

Regarding accuracy, Turkish is again the language with the lowest accuracy in nearly all models, with the exception of Llama, where it attains an accuracy of 38.50%. In contrast, English achieves the highest accuracy in most models, except in Llama, where the accuracy drops to 36.80%. Dutch generally demonstrates intermediate accuracy, ranking neither highest nor lowest across all seven models. Spanish follows a pattern similar to Dutch, showing moderate accuracy in six models and achieving the highest accuracy in Llama, emulating its performance in an ambiguous context. Notably, for every language except for Dutch, Llama stands out as the only model in which the performance in accuracy deviates from the pattern in the other models.

To better illustrate the relationship between bias scores and accuracy we created a scatterplot shown in Figure 2. It reveals that bias scores in the ambiguous context are not only higher but also more variable across all accuracy levels and languages, compared to the disambiguated context. In the disambiguated context, bias scores are much lower, rarely exceeding 0.04, and remain close to zero in all four languages, indicating greater stability. The scatterplots also reveal no clear positive or negative relationship between accuracy and bias scores in either context. For both disambiguated and ambiguous context, bias scores do not consistently increase or decrease as accuracy changes. Another notable pattern is that in the ambiguous context, Dutch and Spanish exhibit the highest bias peaks regardless of accuracy, while in a disambiguated context, Dutch, Spanish, and Turkish show the highest peaks, independent of accuracy.

Overall, the results indicate consistent and distinguishable patterns of bias and accuracy across languages and contexts, while also demonstrating patterns that are not uniformly distributed. Spanish emerges as the most biased language across models and contexts, while Dutch consistently demonstrates moderate bias levels. In an ambiguous context, English and Turkish exhibit similar varied bias patterns, primarily differing in model performance. In contrast, the disambiguated context reveals greater variability in Turkish bias scores, whereas English displays the lowest and most stable bias. Regarding accuracy in both contexts, English consistently achieves the highest accuracy in both contexts, Dutch maintains moderate accuracy, and Turkish records the lowest

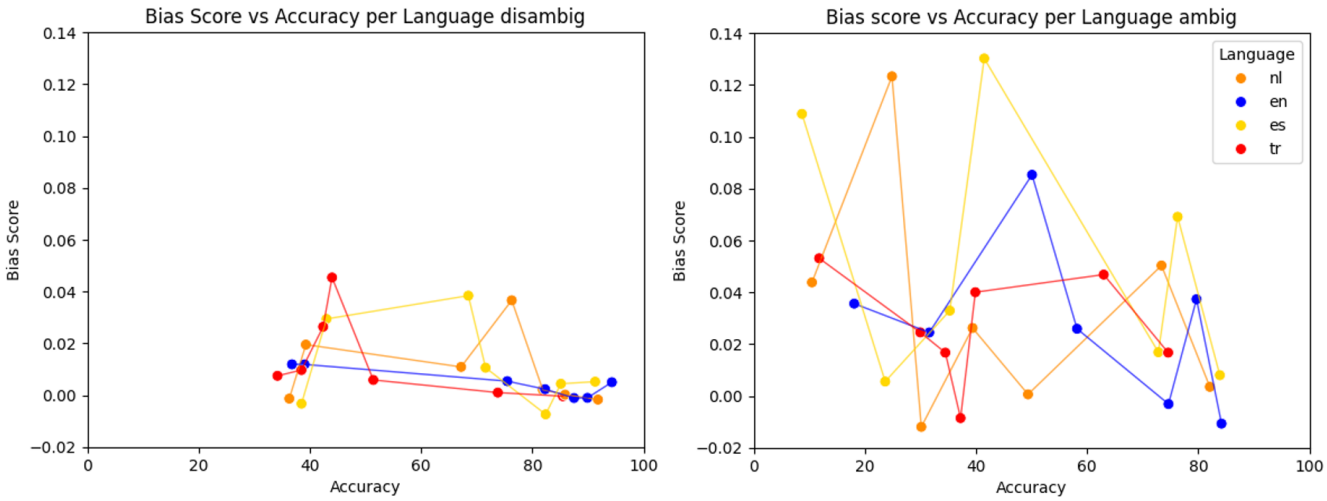

Figure 2: Bias Score vs. Accuracy per language in both contexts

accuracy. Spanish generally follows Dutch with moderate accuracy.

The scatterplot analysis illustrates that bias scores in the ambiguous context are higher and more variable compared to the disambiguated context, where bias values cluster more tightly around zero, indicating greater stability. Importantly, no clear positive or negative correlation between bias and accuracy is observed. Notably, Dutch and Spanish show the highest bias peaks in the ambiguous context, while in a disambiguated context, Dutch, Spanish, and Turkish exhibit the most pronounced bias peaks.

5.2 Model consistency and accuracy

As mentioned in section 4.4, we used the MAD in order to identify the most consistent and most accurate model. Table 4 presents the MAD values for each model in various metrics, where lower values indicate greater consistency, and higher values indicate greater inconsistency. To select a single model, we calculated the overall MAD's across models for bias and accuracy. The Tables 5 and 6 present the MAD values for accuracy and bias across seven models. The MAD values for accuracy range from 2.46% (Qwen2) to 14.28% (Phi 3.5 mini), while MAD values for bias range from 0.00% (Llama) to 0.02% (Zephyr). These results indicate that Qwen2 has the lowest overall MAD for accuracy and Llama has the lowest overall MAD for bias, suggesting that Qwen2 is the most consistent in terms of accuracy and Llama the most consistent when it comes to bias.

To measure accuracy, we used the method described in Section 4. The results are summarized in Tables 7 and 8. Table 7 illustrates that the accuracy in disambiguated context ranges from 37.02% (Qwen2) to 90.73% (Aya), while in the ambiguous context, accuracy spans from 12.28% (Aya) to 81.20% (GPT 3.5). Aya stands out by exhibiting the highest accuracy in disambiguated context, but showing the lowest accuracy in ambiguous context, which demonstrates its struggle with ambiguous questions. Table 8 shows the overall accuracy for the seven models, ranging from 33.51% (Qwen2) to 81.80% (GPT 3.5). Among these, GPT 3.5 is the most accurate model with an accuracy of 81.80%, followed by Mistral with 68.83% and Phi 3.5 mini closely trailing at 67.81% accuracy.

	AccuracyA	AccuracyD	BiasA	BiasD
Aya	35.64	25.21	0.0249	0.0077
GPT 3.5	33.28	16.89	0.0310	0.0116
Llama	12.33	26.11	0.0080	0.0079
Mistral	25.18	9.82	0.0155	0.0105
Phi 3.5 mini	16.45	18.62	0.0340	0.0069
Qwen2	17.93	28.49	0.02	0.0071
Zephyr	9.91	13.42	0.06	0.0199

Table 4: MAD per metric and model

Model	MAD for accuracy
Aya	2.76
GPT 3.5	3.80
Llama	5.27
Mistral	7.67
Phi 3.5 mini	14.28
Qwen2	2.46
Zephyr	9.79

Table 5: MAD for accuracy

Model	MAD for bias
Aya	0.0137
GPT 3.5	0.0054
Llama	0.0048
Mistral	0.0114
Phi 3.5 mini	0.0050
Qwen2	0.0089
Zephyr	0.0219

Table 6: MAD for bias

Model	AccuracyD	AccuracyA
Aya	90.73	12.28
GPT 3.5	82.40	81.20
Llama	39.40	40.73
Mistral	64.58	73.10
Phi 3.5 mini	77.09	58.53
Qwen2	37.02	29.99
Zephyr	67.38	39.10

Table 7: Accuracy per context type

Model	Overall accuracy
Aya	51.51
GPT 3.5	81.80
Llama	40.07
Mistral	68.84
Phi 3.5 mini	67.81
Qwen2	33.51
Zephyr	53.24

Table 8: Overall Accuracy per model

	Accuracy	MAD accuracy	MAD bias	Combined
Aya	-0.30	0.88	-0.56	0.01
GPT 3.5	1.47	0.64	0.77	0.96
Llama	-0.97	0.30	0.86	0.06
Mistral	0.71	-0.25	-0.20	0.09
Phi 3.5 mini	0.65	-1.78	0.83	-0.10
Qwen2	-1.36	0.95	0.20	-0.07
Zephyr	-0.20	-0.75	-1.89	-0.95

Table 9: Z scores of accuracy and the MAD's

To determine which model is overall the most consistent in bias and in accuracy we created a scatterplot, visualized in Figure 3. This figure shows the relation between MAD of bias and the MAD of accuracy presented in a scatterplot.

To determine which model is overall the most consistent in bias and accuracy, we created a scatterplot, visualized in Figure 3. This figure illustrates the relation between MAD of bias and the MAD of accuracy presented in a scatterplot. This figure reveals that Zephyr is the least consistent model in terms of bias, but ranks second to last when it comes to accuracy consistency. Phi 3.5 mini stands out as the least consistent in accuracy out of all models, but is the second best in bias consistency. The two models demonstrating the greatest consistency across metrics are Qwen2 and GPT 3.5. Qwen2 is the most accurate of all models and is the fourth most unbiased model, whereas GPT 3.5 ranks third in consistency in both accuracy and bias.

To eventually determine which model to focus on, we combined the accuracy values with the MAD values of accuracy and bias. These three values have different scales and in order to put these values in the same scale, we used the z-score normalization, where we assign a z-score to each value. This is done by subtracting the values with the mean and dividing them with the standard deviation (17). The mean of the z-scores for the three values is calculated in order to determine the model that is the most accurate and most consistent. This is visualized in Table 9. Here we detect the z scores for Accuracy, MAD accuracy, MAD bias, and the mean of these z score for each model. Among these GPT 3.5 clearly has the highest z-score of 0.96, while the second most consistent model has a combined z-score of 0.09 (Mistral). Zephyr notably has the lowest combined z-score.

However, given this thesis's focus on investigating social bias in LLMs across multiple languages, particularly regarding grammatical gender and cultural masculinity based on Hofstede's model, bias consistency is most relevant. Phi 3,5 mini demonstrates the highest consistency in bias across

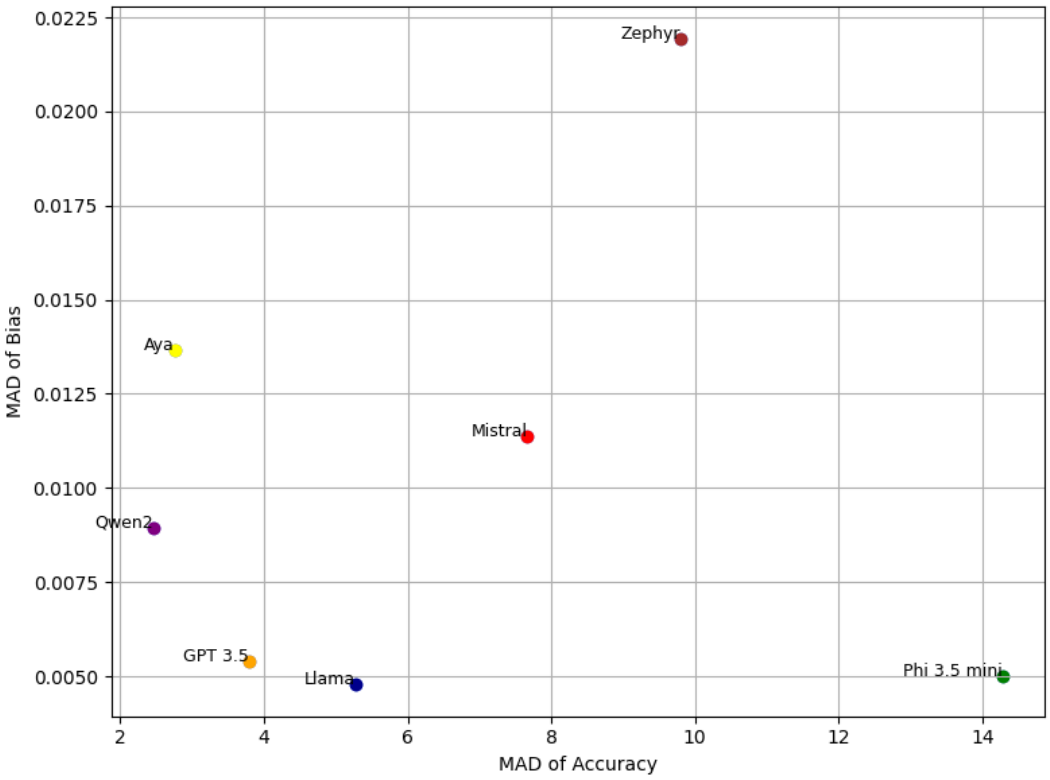

Figure 3: MAD of accuracy vs. MAD of bias

languages and contexts, signifying its suitability for this research focus. Moreover, as described in section 4.1, we have applied the MBBQ benchmark extensively to Phi 3.5 mini, resulting in significantly more data available for this model compared to others. This abundance of data enhances the reliability and stability of our results, supporting its primary role in our analysis. Therefore, due to its high bias consistency and the larger volume of available data, we select Phi 3,5 mini as the primary model for our detailed analysis. However, given our interest in understanding variability across languages and what causes these differences, we also consider and discuss cross-linguistic patterns and potential contributing factors, ensuring a comprehensive investigation beyond mere model consistency.

5.3 Bias Score and Linguistic Gender Category

Now we investigate the relationship between bias scores and the linguistic gender categories in Phi 3.5 mini, using the method described in 4.5.1, and visualized in Figure 4 and Figure 5.

In Figure 4 we can observe that the higher the grammatical gender index, the lower the bias score in both disambiguated and ambiguous context, but there are a few exceptions. In disambiguated context, Turkish does not follow this pattern, Turkish shows the highest bias score while being the least gendered language of the group (index 5). In an ambiguous context, the bias score first increases from index 1 to 2, but then the bias score decreases, breaking the general observed pattern. What is notable is that Phi 3.5 mini in disambiguated context has more positive bias scores and in the ambiguous contexts has more negative scores, this suggests that the model exhibits more biased

	Bias score disambig	Bias score ambig
Dutch	0.0021	0.0006
Spanish	0.0044	0.0169

Table 10: Bias scores for gendered languages

	Bias score disambig	Bias score ambig
English	-0.0010	-0.0031
Turkish	0.0059	-0.0086

Table 11: Bias scores for genderless languages

responses in disambiguated context and more counter-biased responses in ambiguous context.

In Figure 5 we detect that the subsets Age, Physical appearance, and SES exhibit similar a patterns. In both contexts, the bias scores of these subsets increase slightly as the grammatical gender index rises, which is contrary to the overall trend that we have detected in the previous figure. What is notable is that the subsets Gender identity and sexual orientation have sharp ups and downs. In an ambiguous context, sexual orientation decreases sharply from index 1 to 2, while gender identity increases sharply from 2 to 3. In a disambiguated context, sexual orientation decreases sharply between indices 2 and 4, then rebounds between 3 and 5, and gender identity increases sharply from index 2 to 3. Additionally, in both contexts, sexual orientation bias scores tend to decrease as the grammatical gender index increases, followed by an increase from index 3 to 5, mirroring the pattern seen in Figure 4 for the disambiguated context. Overall, these scatter plots suggest that the relationship between bias cores and grammatical gender index is complex and more nuanced rather than straightforward. However, it is clear that different subsets can distinctly influence bias scores.

Using the two-sampled t-test across all bias scores and the gendered and genderless categories, the t-test statistic and the p values were calculated, and visualized in Table 12. The value in both contexts are smaller than 0.05, suggesting that the difference in mean bias scores between the two categories is statistically significant in both contexts. The calculated t-test statistics in disambiguated and ambiguous context, are -4.186 and -4.193. Since we used gendered as group A and genderless as group B (gendered - genderless), these negative t-statistics indicate that genderless languages exhibit higher bias scores than gendered languages in both contexts.

Tables 13 and 14 display the results of the linear regression in disambiguated and ambiguous context predicting bias scores from grammatical gender indices. In both ambiguous and disambiguated context, the linear regression shows that genderless languages, with a higher gender index, have a higher mean bias score compared to gendered languages, that have a lower gender index, by 0.005 units in ambiguous context and by 0.007 units in disambiguated context. The difference is

	T-test statistic	P-value
Disambiguated	-4.186	2.836e-05
Ambiguous	-4.193	2.762e-05

Table 12: Two-sample t-test of the bias scores in gendered and in genderless categories in both disambiguated and ambiguous contexts

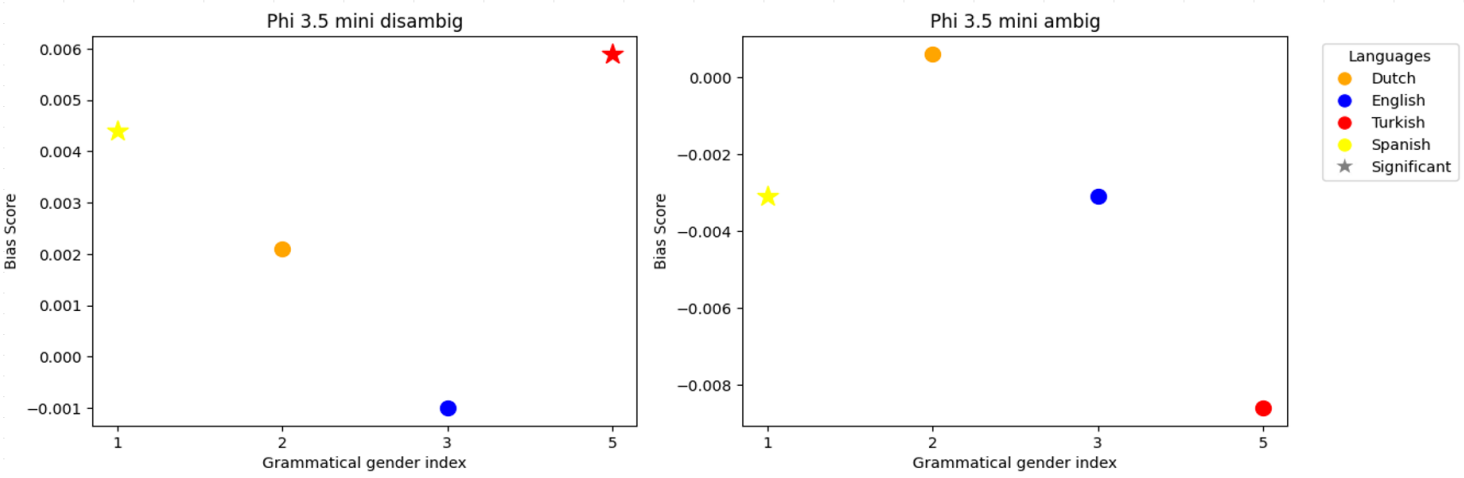

Figure 4: Overall bias scores across grammatical gender indices in both contexts for Phi 3.5 mini. The orange, blue, yellow and red dots/ stars represent the languages Dutch, English, Spanish, and Turkish. The bias score that are marked with a star are significantly different from 0 ($p < 0.05$ using the Wilcoxon Signed-Rank Test.)

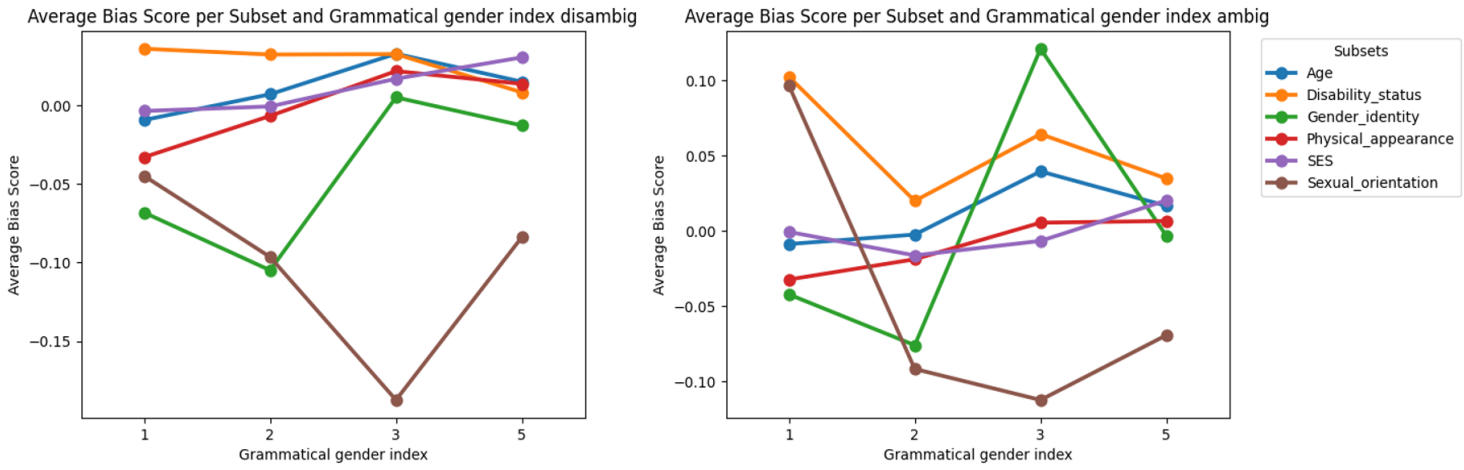

Figure 5: Bias scores per subset vs. Grammatical gender index in both contexts for Phi 3.5 mini.

	coef	std err	t	P>|t|
const	-0.011	0.006	-1.926	0.054
Grammatical gender index	0.007	0.002	3.484	< 0.001

Table 13: Linear regression assessing the relationship between bias scores and Grammatical gender index in disambiguated context. The "grammatical gender index" coefficient represents the estimated mean difference in bias scores in different grammatical gender indices. P-values were obtained by testing the significance of each coefficient. The significance is indicated by < 0.001.

	coef	std err	t	P>|t|
const	-0.006	0.006	-0.957	-0.957
Grammatical gender index	0.005	0.002	2.603	0.009

Table 14: Linear regression assessing the relationship between bias scores and grammatical gender index in ambiguous context. The significance is indicated by 0.009.

significant as the p values in both contexts are < 0.05. This suggest that the grammatical gender index has a meaningful impact on bias scores in both contexts. Overall it can be stated that the results consistently indicate that there is a significant association between bias score and the categories Genderless and Gendered in both contexts. To summarize, we can state that linguistic gender category is an important factor influencing bias in LLMs, but the relationship is multifaceted and influenced by specific social subsets and contextual factors.

5.4 Bias Score vs. Masculinity Score

In this section we move on to investigate the bias scores vs. the masculinity scores (MAS) in the most consistent and accurate model, Phi 3.5 mini. In order to investigate the relationship between these variables, we created a scatterplot as visualized in Figure 6 and in Figure 7.

In Figure 6 we can observe that Spanish has one of the highest bias scores in both contexts and has the second lowest MAS score. English however has the highest MAS score and has one of the lowest bias scores across all languages and contexts. Dutch stays in the middle with the bias scores, even though it has the lowest MAS score of all the languages. Turkish is most noteworthy due to having a drastic difference in bias scores across different contexts, with a bias score of 0.0059 in disambiguated context, and a bias score of -0.0086 in ambiguous context.

Figure 7 displays the relationship between the bias scores and the MAS scores per subset. In the left plot, which is the relationship between the Bias scores and the MAS scores in disambiguated context, we can observe that SES, Age, and Physical appearance are close to zero and show little to no fluctuations. The subject sexual orientation is notably the most biased in this context, with bias scores fluctuating between 0.1000 and -0.1000. With gender identity, we can clearly observe that there is a positive correlation between bias scores and MAS scores.

In ambiguous context, the right plot, we can again detect that SES, Age, and Physical appearance are close to zero and show little to no fluctuations. In this context Disability status is more similar to SES, AGE and Physical appearance, in contrast to Disability status in disambiguated context, where it is further away from zero and has more fluctuations. Again we see here that the gender identity has a positive correlation between bias scores and MAS scores. Sexual orientation is notably

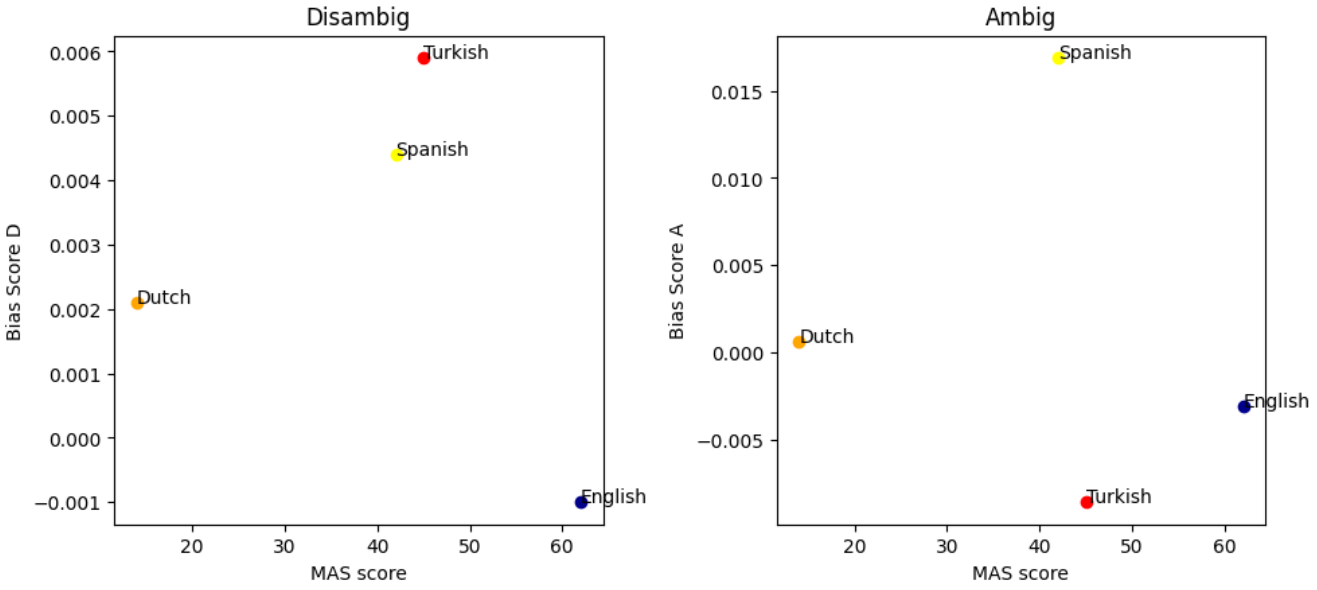

Figure 6: Bias score vs. MAS score

consistently negative and decreases sharply with higher MAS scores.

In both contexts it is notable that disability status has the highest positive scores, especially at lower MAS, and it is also notable that sexual orientation in both contexts show the strongest negative bias.

Using the Pearson Correlation Coefficient across all bias scores and MAS scores, as mentioned in 4.5.2, we calculated the correlation scores. These scores are 0.0093 for bias in disambiguated context, and 0.0159 for bias in ambiguous context. These values are approximately zero, indicating no significant relationship between these variables. This suggests that MAS scores do not predict bias scores in the dataset, and any association is likely negligible or non-existent according to this test.

Despite the previous results that indicate no association we have carried out a linear regression across all bias scores and MAS scores, the results are presented in the Tables 15 and 16. The linear regression analysis in disambiguated context reveals a significant but very small positive association between MAS scores and bias scores, with a coefficient of 0.001, and a p-value of 0.005. This infers that as MAS scores increase, bias scores tend to increase very slightly, 0.001 for each one-unit increase in MAS score. In ambiguous context, linear regression also shows a statistically significant positive relationship that is very minimal. With a p-value of < 0.001, making it highly statistically significant, and a coefficient of 0.001, which is a very minimal increase per one-unit MAS score increase. These linear regressions in both contexts indicate a highly significant but very small positive relationship.

Overall, while the Pearson Correlation Coefficient indicates that MAS scores do not strongly predict bias scores, the linear regression analyses reveal subtle positive relationships. Additionally, bias varies meaningfully across different languages and social subsets.

	coef	std err	t	P>|t|
const	-0.012	0.007	-1.673	0.094
MAS score	0.001	< 0.001	2.819	0.005

Table 15: Linear regression examining the relationship between bias scores (aggregated across all bias categories and languages) and Hofstede's masculinity vs. femininity (MAS) scores in the disambiguated context. The coefficient for "MAS score" represents the estimated change in bias score per one-unit increase in MAS. Reported are the unstandardized coefficients (*coef*), their standard errors (*std err*), t-statistics (t), and p-values ($P>|t|$).

	coef	std err	t	P>|t|
const	-0.023	0.007	-3.222	0.001
MAS score	0.001	< 0.001	4.707	< 0.001

Table 16: Linear regression examining the relationship between bias scores and MAS scores in the ambiguous context.

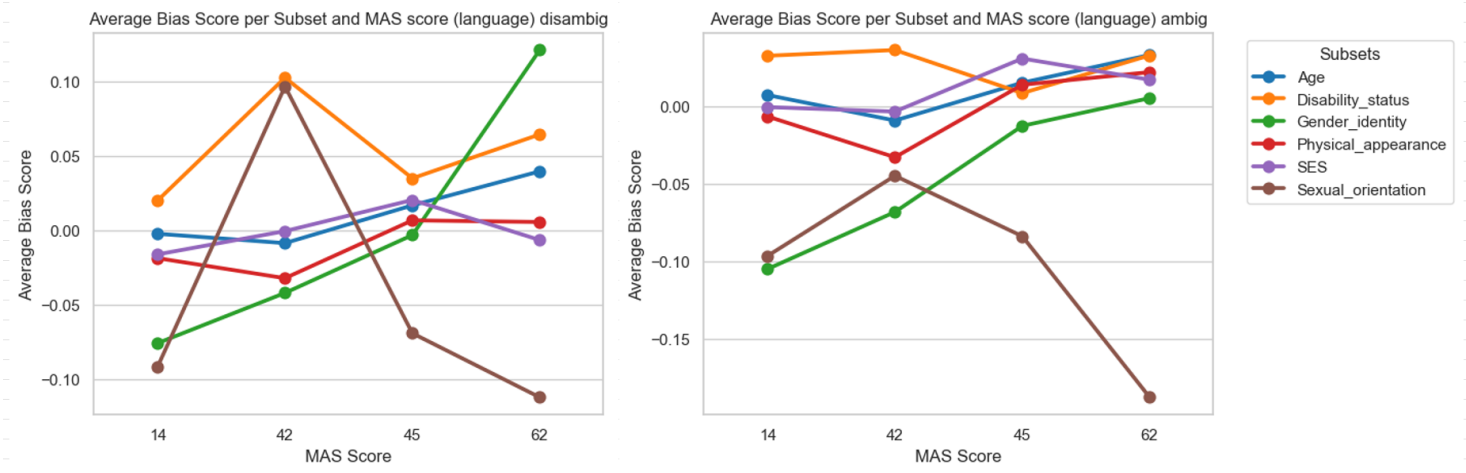

Figure 7: MAS score vs. bias score per subset (bias category)

6 Discussion

This study aimed to explore how social bias in LLMs varies across languages, and to what extent these differences can be explained by grammatical gender and Hofstede's masculinity vs. femininity cultural dimension. By analyzing bias scores across four languages with distinct linguistic gender profiles and MAS scores, and by examining in both a disambiguated and a ambiguous context, our findings reveal a nuanced and diverse relationship between language, culture, model, and bias.

6.1 Findings

Looking at all seven models, our results demonstrate consistent and distinguishable patterns of social bias and accuracy across languages and contexts. Certain languages, notably Spanish and Dutch, exhibit characteristic bias profiles, and regarding accuracy, all languages exhibit characteristic profiles. The scatterplot analysis confirms a persistent lack of correlation between bias and accuracy and further reinforces the recurring bias patterns observed for Spanish and Dutch. However, these patterns are not uniformly distributed, the nature and extent of bias varies depending on the specific language, the contextual setting, and the underlying LLM model. For instance, Spanish consistently emerges as the most biased language across models and context, whereas Dutch consistently demonstrates moderate bias levels, indicating that bias levels depend on language. The scatterplot shows that bias scores in the disambiguated context are considerably lower and less variable compared to the ambiguous context, where bias scores are higher and more dispersed, highlighting the distinct bias distributions by context. This pattern can be explained by the reduced availability of contextual cues for the right answer, which forces models to rely more heavily on their own biases(48; 33). English and Turkish display similar, varied bias patterns in an ambiguous context but differ in model-specific performance, pointing to the influence of the underlying LLM model used on bias manifestation. These findings, obtained using the MBBQ benchmark, suggest that certain languages may be more susceptible to social bias in LLMs, and that the manifestation of bias and accuracy can shift depending on both context, in which questions are asked, and the specific model considered. This underscores the importance of examining bias within a multilingual and multi-model framework to comprehensively understand the dynamics of social bias in LLMs.

When focusing exclusively on Phi 3.5 mini, both similarities and differences emerge in comparison to the broader set of seven models. The accuracy patterns in Phi 3.5 mini largely reflect the overall trends. English consistently achieves the highest accuracy, Dutch demonstrates moderate accuracy, and Turkish exhibits the lowest accuracy. However, notable differences arise in the bias patterns. For instance, Dutch generally shows moderate bias across all models, but within Phi 3.5 mini in ambiguous context, Dutch displays the lowest bias among the languages. Similarly, while Spanish typically exhibits the highest overall bias, in the disambiguated context for Phi 3.5 mini, its bias score is only moderate. These observations suggest that Phi 3.5 mini is more sensitive to contextual variations in bias and may capture more nuanced shifts in bias patterns across different settings. Importantly, these findings underscore the critical role of question formulation and clarity in influencing bias manifestation in LLMs, highlighting prompt engineering as a valuable approach to mitigate bias.

A key focus in this study was the relationship between bias scores and grammatical gender index. The scatter plots in Figures 4 and 5 reveal a complex and nuanced relationship between bias scores and the grammatical gender index. Generally, higher grammatical gender indices are

associated with lower bias scores in both ambiguous and disambiguated contexts, although there are exceptions. For instance, Turkish, the least gendered language of the group, shows the highest bias score in the disambiguated context, breaking the general pattern. Furthermore Phi 3.5 mini exhibits more positive bias in disambiguated contexts and more negative bias in ambiguous contexts, further underscoring context-dependent bias scores. Subsets such as age, physical appearance, and SES show slight increases in bias scores with increasing gender index, contrary to the overall trend, while gender identity and sexual orientation display sharp fluctuations, with sexual orientation bias generally decreasing as gender index increases before rebounding at higher indices. These patterns indicate that the relationship between linguistic gender and bias is not straightforward and that the different social subsets distinctly influence bias scores.

Statistical testing revealed that in both ambiguous and disambiguated context, there is a significant association, with the genderless category exhibiting higher bias scores than the gendered category. This pattern is further supported by linear regression analyses, which showed significant differences between genderless and gendered categories in both contexts. These findings suggest that grammatical gender is relevant to bias expression in LLMs. One possible explanation is when fewer linguistic gender cues are present, such as in genderless languages (16), the LLM has fewer explicit information to process, which will lead the LLM to rely more heavily on patterns and stereotypes present in its training data, as mentioned before (48; 33).

Another key focus in this study was the relationship between bias scores and MAS scores. Visual inspection of the scatter plot revealed that MAS does not have a simple direct correlation with bias scores. Spanish with one of the lowest MAS values, exhibits some of the highest bias scores in both ambiguous and disambiguated contexts. Conversely, English which has the highest MAS, shows among the lowest bias scores. Dutch, despite having the lowest MAS, maintains moderate bias scores, indicating how low masculinity does not necessarily correspond to low bias. Turkish presents a unique case, showing a significant shift in bias between contexts, highlighting how contextual factors may interact with cultural dimensions and influence bias. These patterns suggest that, while cultural masculinity might play a role in shaping bias, it is not the sole determinant. Other linguistic, contextual, and model-specific factors likely contribute to the observed variations in bias across languages.

A closer look at bias scores by subset reveals further complexity in how bias manifests across different groups, as illustrated in Figure 7. Bias is clearly not uniform across bias categories. For instance, sexual orientation exhibits the most negative scores in both disambiguated and ambiguous context of all bias groups. This negative trend becomes even more prominent as MAS score increases. In contrast disability status constantly shows the highest positive bias scores, especially at lower MAS scores. Gender identity stands out in disambiguated context, showing a sharp rise in bias score with the highest MAS score, inferring a strong interaction between language, context, and bias category. Meanwhile SES, age and physical appearance generally remain close to zero, displaying minimal bias and little fluctuation regardless of MAS score or context type. These patterns underscore the importance of dissecting the bias analysis by bias category. The effect of MAS score on bias is dependent on the subsets, with some categories showing an increased bias score with higher MAS, such as age, gender identity, physical appearance, and SES, and other categories showing the opposite or no clear pattern. This finding aligns with research done by Ghai et al. (13), 2021 and Neplenbroek et al. (33), 2024, emphasizing the need for more intersectional analyses of bias in AI systems.

To quantitatively assess these relationships, Pearson Correlation coefficients were calculated

across all bias and MAS scores, yielding values near zero (0.0093 for disambiguated and 0.0159 for ambiguous contexts), indicating no significant correlation. However, linear regression analyses revealed statistically significant but very small positive associations between MAS and bias scores in both contexts ($p < 0.001$ in both contexts), suggesting that bias tends to increase slightly with MAS, though the effect is very minimal.

Finally, to answer our research question:

> "How does social bias in LLMs vary across languages, and to what extent can these differences be explained by grammatical gender and cultural masculinity as defined by Hofstede's model?"

We can now conclude that social bias in LLMs varies significantly across languages, contexts, and bias categories. Part of this variation can be explained by grammatical gender, where genderless languages tend to show higher bias, likely because models compensate for missing gender cues by relying on their own biases. Hodstede's masculinity vs. femininity (MAS) dimension only has a weak and complex relationship with bias, meaning cultural dimensions alone cannot account for the differences. The underlying LLM used also matters, as some models, like Phi 3.5 mini are more sensitive to contextual changes. Bias patterns are also category specific, with some stereotypes, such as age, gender identity, physical appearance, and SES behaving differently across MAS levels and languages. Finally prompt contexts, in which questions are formulated and asked, strongly influences bias, as ambiguous contexts produce both higher and more varied bias scores compared to disambiguated context, as models have to depend more on stereotypes.

Our results align closely with the main conclusion reported by Neplenbroek et al. (33), 2024, in their evaluation of multilingual bias in using the MBBQ benchmark. As in their study, we observe that social bias and accuracy in LLMs differ significantly across languages and contexts, with Spanish consistently emerging as the language most susceptible to bias. Both studies confirm that models tend to display higher bias scores in ambiguous contexts compared to disambiguated contexts. We also find, consistent with their results, that bias varies substantially across different bias categories, underscoring the need for category-wise analysis rather than aggregate bias scores.

Importantly, by incorporating new models such as, Qwen2 and Phi 3.5 mini, and integrating linguistic gender and Hofstede's masculinity vs. femininity cultural dimension into our analysis, our study extends Neplenbroek et al. (33)'s cross-lingual work. We demonstrate that certain models, such as Phi 3.5 mini, exhibit heightened sensitivity to contextual variations in bias, revealing more nuanced shifts across languages and contexts. Additionally, our inclusion of grammatical gender and MAS scores provides deeper insight into the linguistic and cultural factors mediating bias expression, aspects less emphasized in Neplenbroek et al. (33)'s initial evaluation. While our findings largely corroborate Neplenbroek et al. (33)'s conclusions, subtle do divergences emerge, such as unique context-sensitive bias patterns in Turkish, including the highest bias in some contexts. These differences are likely attributable to differences in model selection and have not been reported in Neplenbroek et al. (33), 2024.

6.2 Limitations

This research has several important limitations. Firstly, the analysis was performed in only four languages. This restricted the generalizability of the findings and limited the statistical power to detect robust cross-cultural trends.

Another limitation is that our research only focused on grammatical gender, MAS scores, different models, different contexts, and different languages. We did not take into account other cultural factors, such as individualism and history with immigration and discrimination, or technical factors, such as model architecture or tokenization, which could influence bias in LLMs.

The results are further shaped by model constraints, as most analyses were performed exclusively on Phi 3.5 mini. While this model demonstrated the highest consistency in bias, it was the least consistent in accuracy and had the lowest combined z-score among the evaluated models. This exclusive focus may have limited the overall robustness and balance of the findings, since other models, such as Qwen2 and GPT 3.5, exhibit higher consistency in both bias and accuracy along with higher combined z-scores, which could potentially reveal different or more nuanced bias patterns. By not analyzing using these models, the study may have missed hidden bias patterns and opportunities for more reliable results. Future research incorporating a broader range of models with strong performance across multiple metrics would likely yield a more comprehensive understanding of bias and improve the generalizability and accuracy of conclusions.

Finally, there was an imbalance in the number of data points across bias categories (subsets), with some subsets being represented more heavily than others. This may affect the reliability and interoperability of the observed patterns.

6.3 Future work

A suggestion for future research could be to further expand the linguistic and cultural diversity of studies on bias in LLMs by incorporating a wider range of languages and cultural contexts. This could enhance the generizability of the findings. Exploring additional cultural and linguistic variables, such as Hofstede's power distance dimension or individualism vs. collectivism dimension, and language families may provide further insight into how underlying bias works, beyond what is captured by grammatical gender and MAS alone.

Another suggestion could be to do qualitative research on why certain bias categories, such as sexual orientation, exhibit higher bias scores than others, as seen in Figure 7. This work could help clarify how different bias categories shape biased outcomes in LLMs. Complementing the research with quantitative approaches with in depth qualitative studies could further illustrate why certain categories are particularly susceptible for bias.

Another potential avenue for future research is to investigate how variations in prompt formulation influence the manifestation of bias. AS demonstrated in Section 6.1, different contexts produced markedly different bias levels. It would be particularly valuable to examine the extent of prompt modification and the specific linguistic changes required, such as word choice or phrasing needed to elicit measurable shifts in bias.

Additionally, research on how real world events in social attitudes are reflected in model outputs over time could provide valuable insights into the dynamic interplay between AI systems and the cultures in which they are deployed. A method would be to use cross-sectional analyses to help inform strategies for ongoing monitoring and mitigation of bias in AI systems.

7 Conclusions

This paper looked at how social bias in LLMs varies across languages, and to what extent these differences can be explained by grammatical gender and hofstede's masculinity vs. femininity cultural dimension. Our results found that social bias in LLMs demonstrate consistent and distinguishable patterns of social bias, but do not perform uniform across languages. The bias patterns can be explained by grammatical gender can not be explained by cultural masculinity measured with MAS by Hofstede. The relationship between language gender characteristics, cultural masculinity and bias is more nuanced, context-depended, and varies across bias categories. Nevertheless, through this study we aim to contribute to the ongoing efforts in multilingual debiasing, with the long-term aspiration of developing models that remain free from bias regardless of the prompted message. We also hope our findings will stimulate further research in this important field.

References

[1] Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone. https://arxiv.org/abs/2404.14219.

[2] Agresti, A. (2018). *Statistical Methods for the Social Sciences*, chapter 7.3, pages 199–201. Pearson, Harlow, 5th edition.

[3] Beatty-Martínez, A. L. and Dussias, P. E. (2019). Revisiting masculine and feminine grammatical gender in Spanish: linguistic, psycholinguistic, and neurolinguistic evidence. *Frontiers in Psychology*, 10.

[4] Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, volume 2 of *Springer Topics in Signal Processing*, pages 1–4. Springer, Berlin, Heidelberg.

[5] Beugelsdijk, S. and Welzel, C. (2018). Dimensions and Dynamics of National Culture: Synthesizing Hofstede with Inglehart. *Journal of Cross-Cultural Psychology*, 49(10):1469–1505.

[Bloomington] Bloomington, I. U. Turkish. Accessed: 02-06-2025.

[7] Casado, A., Palma, A., and Paolieri, D. (2021). The scope of grammatical gender in spanish: Transference to the conceptual level. *Acta Psychologica*, 218:103361.

[8] Chan, M.-Y. and Wong, S.-M. (2024). A Comparative Analysis to Evaluate Bias and Fairness Across Large Language Models with Benchmarks. https://doi.org/10.31219/osf.io/mc762.

[9] Chen, R. (2024). Introducing APIs for GPT-3.5 Turbo and Whisper. https://openai.com/index/introducing-chatgpt-and-whisper-apis/. Accessed: 03-06-2025.

[10] Cui, M., Gao, P., Liu, W., Luan, J., and Wang, B. (2025). Multilingual machine translation with open large language models at practical scale: An empirical study. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

[11] Ding, Y., Zhao, J., Jia, C., Wang, Y., Qian, Z., Chen, W., and Yue, X. (2025). Gender bias in large language models across multiple languages: A case study of ChatGPT. In Cao, T., Das, A., Kumarage, T., Wan, Y., Krishna, S., Mehrabi, N., Dhamala, J., Ramakrishna, A., Galystan, A., Kumar, A., Gupta, R., and Chang, K.-W., editors, *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 552–579, Albuquerque, New Mexico. Association for Computational Linguistics.

[12] Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley & Sons, New York, 2nd revised and enlarged edition edition.

[13] Ghai, B., Hoque, M. N., and Mueller, K. (2021). Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

[14] Google, C. (n.d.). Google Colab. https://colab.google/. Accessed: 11-05-2025.

[15] Grigoryeva, L. and Nazmieva, E. (2023). The gender aspect of the English language. Should we teach gender neutrality in the classroom? *Journal of Education Culture and Society*, 14(2):144–153.

[16] Gygax, P. M., Elmiger, D., Zufferey, S., Garnham, A., Sczesny, S., Von Stockhausen, L., Braun, F., and Oakhill, J. (2019). A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, 10.

[17] Henderi, H., Wahyuningsih, T., and Rahwanto, E. (2021). Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *International Journal of Informatics and Information Systems*, 4(1):13–20.

[18] Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in context. *Online Readings in Psychology and Culture*, 2(1).

[19] Hofstede, G., Hofstede, G. J., and Minkov, M. (2010). *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. McGraw-Hill.

[20] Huang, Y., Zhang, Q., Y, P. S., and Sun, L. (2023). Trustgpt: A benchmark for trustworthy and responsible large language models. https://arxiv.org/abs/2306.11507.

[21] James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python*, pages 69–134. Springer.

[22] Jayaseelan, N. (2023). NVIDIA L4 vs. A100 GPUs: Choosing the right option for your AI needs. https://www.e2enetworks.com/blog/nvidia-l4-vs-a100-gpus-choosing-the-right-option-for-your-ai-needs.

[23] Jeknić, R. (2013). Young women and culture: "masculinity" / "femininity" as cultural dimensions in Geert Hofstede's model of "national culture". https://bib.irb.hr/prikazi-rad?rad=664105.

[24] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

[25] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in One-Criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

[26] Kumar, R. (2025). Top 10 English speaking countries in the world.

[27] Li, X., Chen, Z., M. Zhang, J., Lou, Y., Li, T., Sun, W., Liu, Y., and Liu, X. (2024). Benchmarking Bias in Large Language Models during Role-Playing. https://arxiv.org/html/2411.00585v1bib.bib15.

[28] Lin, X. and Li, L. (2025). Implicit bias in llms: A survey. https://arxiv.org/abs/2503.02776.

[29] Liu, Z. (2023). Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

[30] Masoud, R., Liu, Z., Ferianc, M., Treleaven, P. C., and Rodrigues, M. R. (2025). Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

[31] Mihaylov, V. and Shtedritski, A. (2024). What an Elegant Bridge: Multilingual LLMs are Biased Similarly in Different Languages. Technical report, Oxford Artificial Intelligence Society, University of Oxford.

[32] Mishra, P., Singh, U., Pandey, C., Mishra, P., and Pandey, G. (2019). Application of student's t-test, analysis of variance, and covariance. *Annals of Cardiac Anaesthesia*, 22(4):407.

[33] Neplenbroek, V., Bisazza, A., and Fernández, R. (2024). MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. In *First Conference on Language Modeling*.

[34] Nie, S., Fromm, M., Welch, C., Görge, R., Karimi, A., Plepi, J., Mowmita, N., Flores-Herr, N., Ali, M., and Flek, L. (2024). Do multilingual large language models mitigate stereotype bias? In Prabhakaran, V., Dev, S., Benotti, L., Hershcovich, D., Cabello, L., Cao, Y., Adebara, I., and Zhou, L., editors, *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 65–83, Bangkok, Thailand. Association for Computational Linguistics.

[35] Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. (2022). BBQ: A hand-built bias benchmark for question answering. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

[36] Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., and Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, 6(1):101118.

[37] Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. *ACM eBooks*, pages 469–481.

[38] Ranjan, R., Gupta, S., and Singh, S. N. (2024). A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions. https://arxiv.org/abs/2409.16430.

[39] Rigouts Terryn, A., de Lhoneux, M., KU Leuven, C. f. C. L. C., and KU Leuven, D. o. C. S. (2024). Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French. Technical report.

[40] Rithika, R. (2024). Recent advances in large language models: an upshot. *International Journal of Research Publication and Reviews*, 5(6):137–143.

[41] Schut, L., Gal, Y., and Farquhar, S. (2025). Do multilingual LLMs think in English? https://arxiv.org/html/2502.15603v1.

[42] Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

[43] Taalunie (n.d.). Feiten cijfers - taalunie. https://taalunie.org/informatie/24/feiten-cijfers. Accessed: 01-05-2025.

[44] Terryn, A. R. and De Lhoneux, M. (2024). Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French. https://aclanthology.org/2024.humeval-1.2/.

[45] Thomas, S. (2023). Mean Absolute Deviation (MAD): What It Means and Formula — Outlier.

[46] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A.,

Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288.

[47] Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. (2023). Zephyr: Direct distillation of lm alignment.

[48] Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. (2023). "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. In *Conference on Empirical Methods in Natural Language Processing*.

[49] Wilson, K. and Caliskan, A. (2024). Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *Vol. 7 (2024): Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24)*, 7:1578–1590.

[50] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. (2024). Qwen2 technical report. https://arxiv.org/abs/2407.10671.

[51] Yu, S., Choi, J., and Kim, Y. (2025). Delving into multilingual ethical bias: The msqad with statistical hypothesis tests for large language models. https://arxiv.org/abs/2505.19121.

[52] Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., and Hooker, S. (2024). Aya model: An instruction finetuned open-access multilingual language model.