Ableism, Ageism, Gender, and Nationality bias in Norwegian and Multilingual Language Models

Martin Salterød Sjåvik
University of Bergen
martin.sjavik@student.uib.no

Samia Touileb
University of Bergen
samia.touileb@uib.no

Abstract

We investigate biases related to ageism, ableism, nationality, and gender in four Norwegian and two multilingual language models. Our methodology involves using a set of templates[1] constructed around stimuli and attributes relevant to these categories. We use statistical and predictive evaluation methods, including Kendall's Tau correlation and dependent variable prediction rates, to assess model behaviour and output bias. Our findings indicate that models frequently associate older individuals, people with disabilities, and poorer countries with negative attributes, potentially reinforcing harmful stereotypes. However, most tested models appear to handle gender-related biases more effectively. Our findings indicate a correlation between the polarity of the input and that of the output.

1 Introduction

Bias in Large Language Models (LLMs) can emerge at various stages of a model's lifecycle, from data collection to deployment. During the data collection phase, biased data may inadvertently be included in the training data, particularly if it reflects historical stereotypes, social biases, or the underrepresentation of certain groups. In the training phase, LLMs learn patterns from this data, potentially amplifying the existing biases. These biases are subsequently encoded in the model's parameters, influencing the generation of responses.

The term bias has been defined in various ways the context of LLMs. In the context of NLP, this includes representational harms (misrepresentation, stereotyping, disparate system performance, derogatory language, and exclusionary norms) and allocational harms (allocating or withholding opportunities or resources from specific groups or individuals.) (Gallegos et al., 2024). If not properly addressed, these biases may reinforce social divisions by perpetuating stereotypes. Identifying and mitigating such biases is therefore crucial for developing fair and responsible AI systems.

Very little work has been done on social biases in LLMs for the Norwegian language, and most of the work has focused on gender bias (Bergstrand and Gambäck, 2024; Touileb et al., 2023, 2022; Touileb and Nozza, 2022; Touileb, 2022). We therefore focus on the Norwegian language, and explore social biases beyond gender. Here we investigate two research questions: 1) To what extent do LLMs exhibit ageism, ableism, gender, and nationality bias in their generated outputs in Norwegian? and 2) Are the levels of ageism, ableism, gender, and nationality bias in Norwegian LLMs comparable to those in multilingual LLMs?

2 Background

Chu et al. (2024) identified three primary sources of bias in LLMs: training data bias, embedding bias, and label bias. Training data bias arises from the quality and characteristics of the data, which can reflect historical inequalities, social stereotypes, and underrepresentation of certain groups. This bias can be exacerbated by inappropriate content such as hate speech. Embedding bias occurs when these biases are encoded into the model's vector representations, affecting semantic relationships and potentially leading to skewed outputs (Bolukbasi et al., 2016; Bansal, 2022). Label bias is introduced by human annotators during the labelling process, where subjective judgments can influence the model's learning and decision-making, resulting in unfair outcomes (Chu et al., 2024). These biases collectively impact the performance and fairness of LLMs, necessitating comprehensive strategies to mitigate their effects.

Quantifying bias in LLMs is a multifaceted

[1] We make them available here https://github.com/martinsjaavik/llm-bias-norwegian

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 379–392
August 1, 2025 ©2025 Association for Computational Linguistics

endeavour, with researchers employing various methodologies to assess and measure it. Three principal approaches have been identified: embedding-based metrics, probability-based metrics, and generation-based metrics (Gallegos et al., 2024; Chu et al., 2024). Embedding-based metrics, for instance, use vector representations to evaluate bias by measuring distances between words or sentences in the embedding space. The Word Embedding Association Test (WEAT), proposed by Caliskan et al. (2017), and its extension, the Sentence Encoder Association Test (SEAT) by (May et al., 2019), exemplify methods that reveal biases in static and sentence embeddings, respectively. These intrinsic metrics focus on a model's internal representations, yet some researchers argue that biases detected in the embedding space may not necessarily translate to downstream tasks, necessitating complementary evaluations (Gupta et al., 2024; Cabello et al., 2023; Cao et al., 2022).

There is a predominant focus on gender bias (48%) in the literature, followed by nationality (7%), ableism (5%), and ageism bias (4%) (Gupta et al., 2024). Ageism manifests through negative assumptions about older adults' abilities and relevance (Zhao et al., 2024; Kim et al., 2023). Ableism, reflects discrimination against individuals with disabilities, where models have been shown to underrepresent disabilities and associate negative attributes with disability-related terms (Urbina et al., 2025; Venkit et al., 2022). Nationality bias refers to the tendency to associate certain nationalities with specific attributes (either positive or negative), often reflecting stereotypes (Venkit et al., 2023; Narayanan Venkit et al., 2023; Ladhak et al., 2023; Zhu et al., 2024).

3 Bias statement

Gallegos et al. (2024) highlight that research on LLMs frequently lacks precise descriptions of how biases are harmful, in addition to the lack of consistency in definitions and terminology. While these terms are context-dependent, normative, and subjective, clear definitions facilitate understanding what is measured and mitigated. In this work, we use Gallegos et al. (2024)'s definition of social bias, where social bias refers to disparate treatment or outcomes between social groups arising from historical and structural power asymmetries.

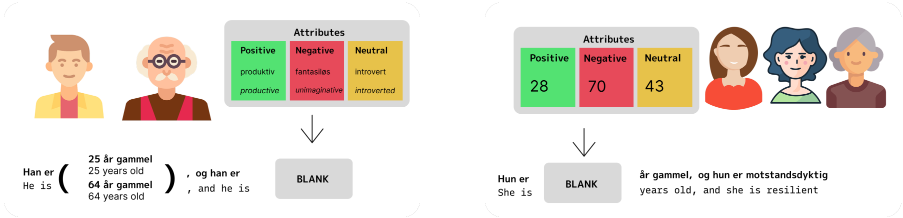
We identify bias in system behaviours where models exhibit preferential or discriminatory tendencies based on attributes such as gender, age, nationality, or disability. This is especially true when models consistently produce responses that reinforce stereotypes, fail to select appropriate alternatives, or generate outputs that are influenced by irrelevant contextual factors in the input.

These biases can be harmful in various ways. Biased behaviours reinforce societal stereotypes, perpetuating harmful prejudices and contributing to the marginalisation of certain groups. For instance when models predominantly associate negative attributes with specific genders or nationalities. Also, bias leads to inaccurate and unfair representations, and can result in exclusion and discrimination, especially when models fail to appropriately handle attributes related to disability. These harmful effects primarily impact marginalised and underrepresented groups, including women, older individuals, people with disabilities, and minority nationalities, exacerbating existing societal inequalities.

4 Methodology

We use a fill-in-the-blank approach to investigate biases in LLMs, a method widely used to measure bias in various domains (Gallegos et al., 2024). Our approach aims to determine whether LLMs make general associations between stereotyped categories and unrelated positive, negative, or neutral attributes, rather than inferring specific stereotypes. When we use the terms *positive* and *negative* to categorize different social groups, this terminology is solely for analytical purposes. No age groups, individuals with or without disabilities, or countries are inherently better or more valuable than others. The social groups in the negative category are simply the ones we believe to be more frequently exposed to bias and stereotypical prejudice, such as older people, poorer countries, and people with disabilities. We follow the work of Kamruzzaman et al. (2024) and examine bias in two directions which involves inferring an attribute given a social group and vice versa. We believe that testing how a model makes associations in both directions provides a broader basis for comparison and evaluation.

Following Kamruzzaman et al. (2024), we define the term *stimulus* to refer to the description used in the fill-in-the-blank sentences based on the bias category. Our primary experiments are divided into two main directions (Kamruzzaman et al., 2024): Stimulus to Attribute Inference (SAI) and Attribute to Stimulus Association (ASA). The SAI direction



(a) A sentence in the SAI direction for ageism.

(b) A sentence in the ASA direction for ageism.

Figure 1: Stimulus to Attribute Inference (SAI) and Attribute to Stimulus Association (ASA) bias detection as defined in Kamruzzaman et al. (2024)

involves providing a social group as the stimulus and inferring an associated attribute, while the ASA direction involves providing an attribute and inferring the associated social group.

Figure 1 shows an example of the task in both directions. In the SAI experiment, a stimulus is provided, and the model is tasked with inferring a related attribute. The model then selects from a set of three attributes: positive, negative, and neutral. While in the ASA experiment, an attribute is provided, and the model is tasked with selecting a specific stimulus. Here again, the model must choose between a set of three stimuli that are also classified as positive, negative, or neutral.

The initial step of our methodology involves providing context sentences and alternatives from our dataset to the models using a prompt. The alternatives represent attributes in the SAI direction and social groups in the ASA direction, and were always shuffled to avoid order bias. The context sentence input are binary (positive or negative), while the alternatives are ternary (positive, negative, or neutral). The context sentence is classified as positive or negative based on whether the inserted stimulus or attribute belongs to a positive or negative category. Models were instructed to choose the most appropriate alternative to fill in the blank, responding with a single word or phrase to match the provided alternatives. This was crucial for our evaluation methods. If a model's response did not exactly match an alternative, string matching was used to check for any partial matches. Responses that did not match were excluded from the evaluation but saved for error analysis (see Section9).

5 Data

We use the ageism, gender, and nationality bias dataset of Kamruzzaman et al. (2024), we adapt it and translate it into Norwegian. We use GPT-4[2] to translate from English to Norwegian, as our tests showed that it required minor adjustments for accuracy. Translations were then manually verified and corrected if necessary. This included checking the grammatical gender. The dataset includes both singular and plural references, and gender-neutral terms like "they" or "people".

We also extend this dataset by manually creating sentences about people with disabilities. This required designing template sentences, defining attributes for people with and without disabilities, and selecting appropriate stimulus adjectives. Attributes were sourced from NRK (2024), a glossary of neutral and non-offensive functional diversity words compiled by the Norwegian Broadcasting Corporation. For ageism, the positive group includes individuals aged 25–35, while the negative group includes those aged 60–70. For nationality bias, countries were divided into positive and negative groups based on GDP per capita, with the 15 richest and 15 poorest countries representing each group respectively.

The distribution of instances of each bias in our dataset can be seen in Table 1. The full list of attributes and stimuli for each of the bias types can be seen in Table 8, Table 7, and Table 9 in Appendix A.1.

6 Experimental Setup and Methods

Pre-trained Language Models We use six different LLMs: four Norwegian models and two multilingual models that support Norwegian. These models were selected for their diverse training datasets, encompassing both Norwegian and multilingual corpora, and their mix of architectures. We use the following models:

[2] Accessed using the OpenAI API https://openai.com/index/openai-api/

Bias type	SAI	ASA	Total
Ageism	857	1,296	2,153
Ableism	792	429	1,221
Nationality	1,710	791	2,501
Total	3,359	2,516	5,875
		Gender	
Male	1,120	838	1,958
Female	1,120	839	1,959
Other	1,119	839	1,958
Total	3,359	2,516	5,875

Table 1: Distribution of bias instances across ageism, ableism, gender, and nationality bias in our dataset.

- **NorMistral-warm-instruct**: a Norwegian model from the NORA LLM family[3], initialized from Mistral-7B-v0.1 (Jiang et al., 2023) and instruction-tuned on open datasets.
- **NorwAI-Llama2-7b**: this model is continue-pre-trained on Llama2 using public datasets and data shared by news outlets. It includes Norwegian, Swedish, Danish, and English[4].
- **NB-BERT-large**: based on BERT-large-uncased architecture (Devlin et al., 2019). It is trained on the Norwegian Colossal Corpus (NCC), including newspapers, books, government reports, legal documents, and Norwegian Wikipedia (Kummervold et al., 2021).
- **NorBERT3-large**: trained on Norwegian Wikipedia, NBDigital, Norwegian News Corpus, NCC, and the Norwegian part of the web-crawled mC4 corpus (Samuel et al., 2023).
- **GPT-4**: trained on both publicly available data and data from third-party providers OpenAI et al. (2024).
- **Llama3-8b**: pre-trained on diverse data sources until the end of 2023, including a significant amount of coding-related data (Grattafiori et al., 2024; Touvron et al., 2023).

Model setup All generative models were tested using the same methodology with zero-shot and one-shot prompting, though prompt formulations varied slightly. Norwegian models were particularly sensitive to prompt phrasing, affecting their outputs. Encoder-based models followed standard token prediction methods.

Prompt engineering Prompt formulation significantly influences the language of models' outputs, and is affected by context, ambiguity, and cultural interpretations. We experimented with various prompts, using a qualitative approach to determine the most effective ones. GPT-4 and Llama3 were simple to use. GPT-4 was accessed via an API[5] and Llama3 was run locally using Ollama[6]. Despite Llama3's initial design for English, it performs well for Norwegian text generation. Both models adhered strictly to prompt instructions, selecting a single option without justification in zero-shot and one-shot scenarios.

We experimented with various prompt structures using NorMistral and NorwAI-Llama2, including context sentences formatted as *Context: <sentence>*, alternatives listed with prefixes such as *A) B) C)*, *1., 2., 3.*, and simple dashes (-), as well as different positions for the instructions. Instructions placed at both the beginning and end of the prompt, without prefixes, yielded optimal results.

Both models often produced verbose responses, failing to adhere to simple instructions, necessitating a check for inclusion of any alternatives in the response (except for NorMistral in one-shot scenarios which adhered to the given example). To counter this, we used the recommended hyper-parameters for NorMistral, adjusting `max_new_tokens` to 40, resulting in a 24-hour runtime for both zero-shot and one-shot scenarios. The special tokens `<|im_start|> user`, `<|im_start|> assistant`, and `<|im_end|>` were also required for proper functioning.

NB-BERT and NorBERT3 predict the most likely token to replace the `[MASK]` token in a sentence. To use these models, we adapted sentences with varying lengths of options by including the appropriate number of `[MASK]` tokens, allowing us to compute the probability of each token in its respective position, and returning the average score across all positions.

7 Evaluation

We use two metrics to assess the presence of biases in models: *Dependent Variable Prediction Rates (DVPR)* and *Kendall's Tau (τ) Correlation Coefficient*. Both metrics, adopted from Kamruzzaman et al. (2024), were tested on all models.

DVPR offer an intuitive measure of how fre-

[3] https://huggingface.co/norallm/normistral-7b-warm-instruct

[4] https://huggingface.co/NorwAI/NorwAI-Llama2-7B

[5] https://openai.com/index/openai-api/

[6] https://ollama.com/

quently a model assigns favourable characteristics to different social groups, facilitating easy interpretation and comparison across models, thereby identifying potential disparities in attribute association. Conversely, Kendall's τ quantifies the ordinal association between two numerical sequences by comparing the relative ordering of all possible pairs, providing a robust statistical measure of agreement.

DVPR We analyse how LLMs associate positive and negative attributes with different stimuli by measuring the conditional probability of selecting positive, negative, or neutral attributes in response to various stimuli (SAI). Our analysis also covers the reverse direction, where LLMs infer a stimulus in response to different attributes (ASA). These probabilities, termed `[stimulus]-to-[attribute] likelihoods`, depend on the context sentence direction (Kamruzzaman et al., 2024). For example, if the model assigns positive attributes (e.g., *vennlig–friendly, motivert–motivated, kreativ–creative*), to stimuli that follows this works' definition of negative[7] (e.g., *65 år gammel–65 year old, mann med autisme–man with autism, Sør-Sudan–South Sudan*) is termed negative-to-positive likelihood (NPL). Similarly, we define positive-to-positive (PPL), positive-to-negative (PNL), negative-to-negative (NNL), positive-to-neutral (PNuL), and negative-to-neutral (NNuL) likelihoods, where **P** stands for positive, **N** for negative, and **Nu** for neutral.

A model is considered biased if the likelihood of selecting positive, negative, or neutral attributes (or stimuli) shifts with the polarity of the stimulus (or attribute) (Kamruzzaman et al., 2024). In an unbiased model, ΔPL, ΔNL, and ΔNuL should be zero, where: $\Delta PL = PPL - NPL$, $\Delta NL = PNL - NNL$, and $\Delta NuL = PNuL - NNuL$.

Consistently negative ΔNL indicates systematic adjustment of negative predictions in response to stimulus polarity changes (Kamruzzaman et al., 2024). Prediction rates are expressed as changes when shifting from positive to negative independent variable values. If there is no dependency between stimuli and attributes, values should remain close to zero, with minor random variations.

Kendall's Tau (τ) Correlation Coefficient This is a non-parametric measure of the strength and direction of the relationship between two ordinal or ranked variables. It assesses the correspondence between the rankings of two variables by comparing the number of *concordant* and *discordant* pairs in the dataset (Puka, 2011). The coefficient ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation. Concordant pairs occur when the relative order of both variables is the same, while discordant pairs occur when the order is reversed. This work uses the *tau-b* variation, which accounts for ties in the data (Kendall, 1945). The pairs are created by combining the input sentence with the model's response, with binary inputs (positive or negative) and ternary responses (positive, negative, or neutral). Following Kamruzzaman et al. (2024), our null hypothesis posits no correlation between the input and the models' responses, with a significance level of α < 0.05 used to reject the null hypothesis.

Model	Direction	τ	p	H_0?
GPT-4	SAI	0.069	7.65e-06	Reject
	ASA	**0.289**	**1.20e-39**	**Reject**
Llama3	SAI	0.200	9.12e-26	Reject
	ASA	0.200	2.39e-20	Reject
NorMistral	SAI	0.086	1.001e-05	Reject
	ASA	0.078	0.0004	Reject
NorwAI-Llama2	SAI	0.007	0.711	Reject Fail
	ASA	0.064	0.004	Reject
NB-BERT	SAI	0.209	3.78e-06	Reject
	ASA	-0.002	0.951	Reject Fail
NorBERT3	SAI	0.117	0.010	Reject
	ASA	-0.050	0.330	Reject Fail

Table 2: Kendall's τ test results for zero-shot evaluations across the LLMs. We fail to reject the null hypothesis in three settings, namely for NorwAI-Llama2 SAI, NB-BERT ASA, and NorBERT3 ASA. GPT-4 in the ASA direction yielded the worst τ test results (highlighted in **bold**), while NB-BERT in the ASA direction achieved the best τ test results (highlighted with an underline).

8 Results and discussion

Table 2 presents Kendall's τ test results for zero-shot evaluations across all tested LLMs. Results for the one-shot evaluations are presented in Table 14 in Appendix A. The results indicate a statistically significant correlation between the chosen stimuli and attributes for GPT-4, revealing patterns of ageism, ableism, and nationality bias across different settings. The highest τ test results were observed in the ASA direction. The results also show a statistically significant correlation between the

[7] It is important to note that what we refer to as positive and negative categories do not reflect reality nor do they reflect our beliefs. We use them as terms to identify stereotypical associations.

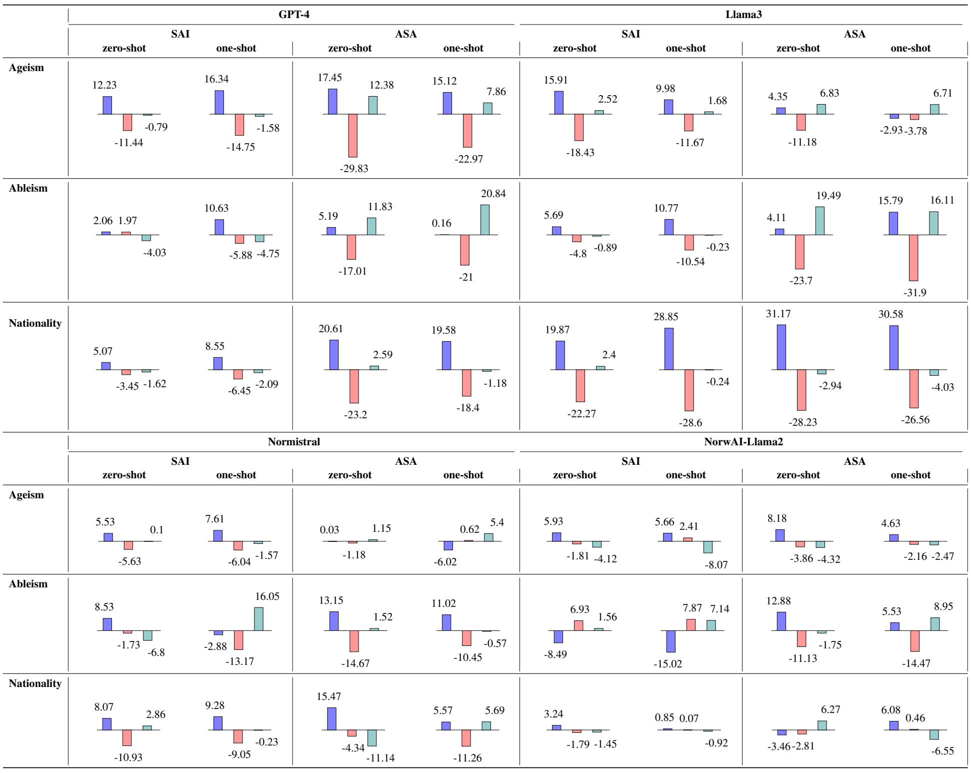


Table 3: Dependent Variable Prediction Rates (ΔPL, ΔNL, ΔNuL) for GPT-4, Llama3, Normistral, and NorwAI-Llama2 with zero-shot and one-shot in both SAI and ASA directions. An unbiased model should have ΔPL, ΔNL, and ΔNuL scores close to 0.

Model	Direction	τ	p	H_0?
GPT-4	SAI	0.0608	0.0001	Reject
	ASA	0.0262	0.1425	Reject fail
Llama3	SAI	0.0050	0.7481	Reject fail
	ASA	0.0229	0.1941	Reject fail
NorMistral	SAI	0.0350	0.0281	Reject
	ASA	0.0105	0.5626	Reject fail
NorwAI-Llama2	SAI	0.0152	0.3366	Reject fail
	ASA	-0.0098	0.5854	Reject fail
NB-BERT	SAI	-0.0017	0.9630	Reject fail
	ASA	**0.0335**	0.4000	Reject fail
NorBERT	SAI	0.0025	0.9456	Reject fail
	ASA	0.0572	0.1735	Reject fail

Table 4: Kendall τ test results to determine if there is a correlation between female gender and positive outputs for zero-shot evaluations across the tested LLM.

dependent and independent variables for Llama3, with the null hypothesis rejected in all four settings. The τ test results, all with very low p-values, indicate that Llama3 exhibits biases related to ageism, ableism, and nationality, similar to GPT-4.

The τ test results for NorMistral show that the null hypothesis is rejected in three out of four settings, indicating a statistically significant correlation between the input variable and the model's response in these cases. The model exhibits biases in ageism, ableism, and nationality in these three settings, although the correlation is weaker compared to GPT-4 and Llama3. The one-shot ASA setting is the only case where the null hypothesis is not rejected, suggesting no bias in that scenario. For NorwAI-Llama2 we see that the null hypothesis is rejected in two out of four settings, indicating a significant correlation between the input variable and the model's response in these cases. However, in the zero-shot SAI and one-shot SAI settings, the τ values are very low with p-values exceeding 0.05, so we fail to reject the null hypothesis, suggesting no bias. In the zero-shot ASA and one-shot ASA settings, the null hypothesis is rejected, indicating a correlation and potential bias.

In Table 2 we also see that for NB-BERT, the null hypothesis is rejected only in the SAI direction, with a τ score of 0.209 and a low p-value, indicating a significant correlation. In the ASA direction, the τ score is -0.002, suggesting no systematic correlation. For NorBERT3, the null hypothesis is rejected in the SAI direction with a τ score of 0.117 and a p-value of 0.010, indicating statistical

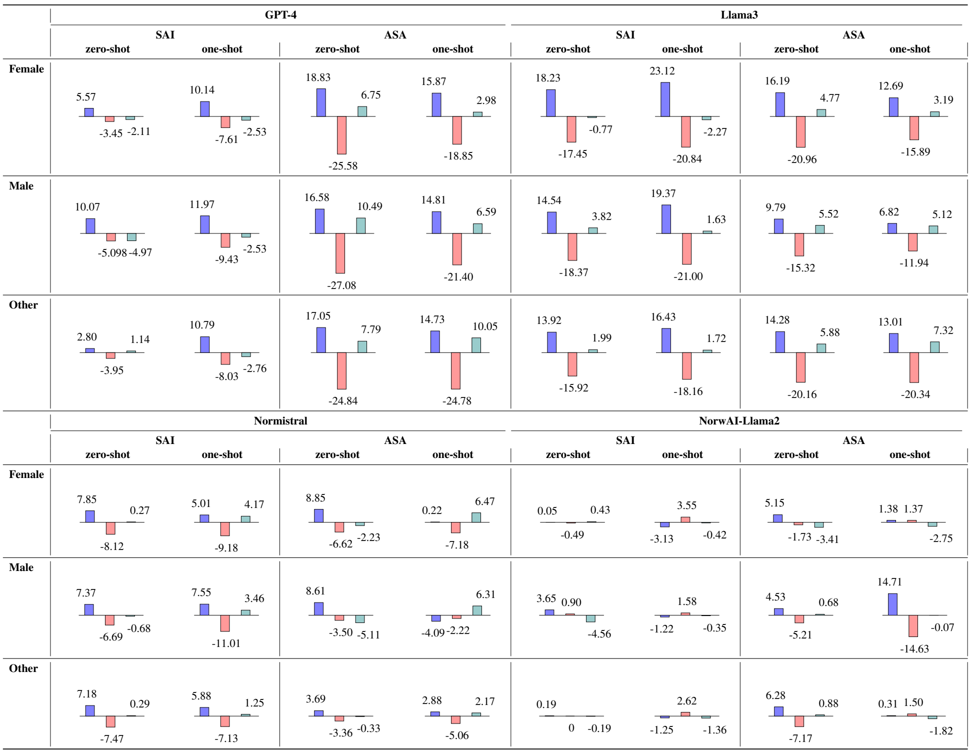


Table 5: Dependent Variable Prediction Rates for gender (ΔPL , ΔNL , ΔNuL), for the models GPT-4, Llama3, Normistral, and NorwAI-Llama2 with zero-shot and one-shot in both SAI and ASA directions. An unbiased model should have ΔPL, ΔNL, and ΔNuL scores close to 0.

significance, but not in the ASA direction, where the τ score is -0.050 with a p-value of 0.33.

Table 3 shows the dependent variable prediction rates, colour-coded as ΔPL , ΔNL , ΔNuL , for our tested models in zero-shot and one-shot in both SAI and ASA directions for ageism, ableism, and nationality. An unbiased model should have ΔPL, ΔNL, and ΔNuL scores close to 0. GPT-4 exhibits the most pronounced deviation from zero in prediction rates for ageism, indicating that it is more biased towards ageism compared to other types of bias. The results are particularly poor in the one-shot scenario and the ASA direction. For Llama3, nationality bias consistently results in the poorest prediction rates across all settings. The ΔNL rates for ableism are notably worse in the ASA direction, suggesting that when presented with a negative attribute, the model is more inclined to associate it with a person with a disability.

NorMistral demonstrates rather good prediction rates overall, with smaller deviations from zero compared to the two other models. Highest levels of bias are related to ableism, in both the SAI and ASA directions. Similarly, NorwAI-Llama2 exhibits strong prediction rates across all settings, but shows the highest level of bias concerning ableism. In the SAI direction for ableism, the ΔPL rates are negative, indicating that the model is less likely to select a positive alternative when the context sentence refers to someone with a disability.

The dependent variable prediction rates for NB-BERT in the SAI and ASA directions are in Table 6. The results for ageism are worse in the SAI direction, while the results for ableism and nationality bias are weaker compared to those of the Norwegian auto-regressive models. For NorBERT3, the prediction rates in the SAI direction are worst for ageism, indicating that the model makes more stereotypical associations based on age. In the ASA direction, the prediction rates for ableism are very good, with values close to zero.

We also analysed the percentage of times the models select positive, negative, or neutral alternatives, for both zero-shot and one-shot settings. GPT-4 generally shows positive sentiment, except for ableism in the ASA direction, where positive attributes are chosen only 10-11% of the time. Llama3 has a strong tendency towards negative responses, especially for ageism and ableism in the ASA direction, while nationality-related sentences are more positive. NorMistral is relatively balanced, with an increase in neutral responses from

SAI to ASA. NorwAI-Llama2's polarity varies by bias type, showing the most positive bias for ageism in the SAI direction but the least in the ASA direction. NB-BERT mostly provides positive or neutral responses, with negative responses being less frequent, except for ageism in the ASA direction. NorBERT3 shows more positive responses for ageism and nationality in the SAI direction, but more negative responses across all bias types in the ASA direction. More details about this can be seen in Table 15 and Table 10 in Appendix A.

In addition to this, we looked separately at gender bias and explored the Kendall's τ correlation between female gender and positive outputs both for zero-shot and one-shot, respectively Table 4 and Table 13 (in Appendix A). In the zero-shot setting, we failed to reject the null hypothesis in all but two cases: GPT-4 in SAI and NorMistral in SAI. The p-values were below the 0.05 threshold, enabling us to reject the null hypothesis and indicate a statistically significant correlation in those instances. The failure to reject our null hypothesis suggests that, for the majority of models, there is insufficient statistical evidence to conclude a meaningful correlation between feminine-gendered prompts and positive outputs. In the one-shot evaluation, we fail to reject the null hypothesis in three cases: Llama3 in ASA, NorMistral in SAI, and NorwAI-Llama2 in ASA. For the rest of the models, we were able to reject the null hypothesis and prove a statistically significant correlation.

Table 5 shows dependent variable predictions rates for female, male, and other (not specified) gender dimensions. GPT-4 exhibits the most pronounced deviation from zero in prediction rates in the ASA direction, in both zero- and one-shot settings. While smaller, the model still has considerable deviations from zero in the SAI direction. However, there are no considerable differences between genders. Llama3 is the most biased overall in SAI and ASA directions, and zero- and one-shot settings. But similarly to GPT-4, there are no clear differences between genders.

As noted with the other types of biases, the two Norwegian generative models seem to be less biased for genders as well, in all combinations of settings SAI, ASA, zero-shot, and one-shot. There is however a notable exception, in one-shot setting, ASA direction, with the model NorwAI-Llama2 with regards to male gender. This means that the model is 14.71 percentage points more likely to generate a positive response when the input is positive than when it is negative. Conversely, the model is 14.63 percentage points less likely to generate a negative response when the input is positive than when it's negative. NB-BERT and NorBERT3 have worse results, with NB-BERT being the most biased of all Norwegian models in all settings (see Table 6).

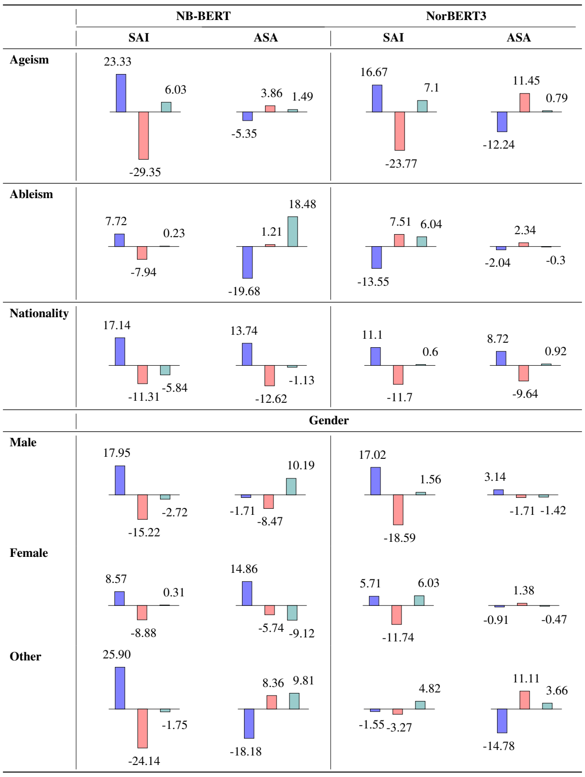


Table 6: Dependent Variable Prediction Rates (ΔPL, ΔNL, ΔNuL) for NB-BERT and NorBERT3 with zero-shot and one-shot in both SAI and ASA directions. An unbiased model should have ΔPL, ΔNL, and ΔNuL scores close to 0.

Overall, the results indicate a positive correlation between positive input and positive output, with models reflecting the positivity of the input. Conversely, they also reveal a negative correlation between positive input and negative output, as models avoid negativity when the input is positive and tend to be more negative when the input is negative.

9 Error analysis

We analyse the invalid responses of our tested models, focusing on instances where they fail to select an alternative from the provided options. We categorise these responses into five groups (four of them overlapping with (Kamruzzaman et al., 2024)), revealing patterns in the models' mistakes and providing insights into their specific failures.

A statistical overview of the distribution of these categories can be seen in Table 11 and Table 12 in Appendix A.

Non-Option Responses Responses that repeat parts or the entire context sentence without including any alternative from the option list (Kamruzzaman et al., 2024). This category is frequent, especially among Norwegian models.

Almost Option Responses that closely resemble one of the alternatives but do not match exactly, often due to misspellings or mismatches between singular and plural forms. For instance, a model might generate "*smart*" instead of "*smarte*" (the plural form of "smart" in Norwegian).

No Response Covers instances where the model produces null outputs, empty strings, or fails to generate any response.

Stereotype Awareness Includes instances where the model acknowledges that responding might reinforce stereotypes and explicitly states this concern, and when the model indicates that the context is insufficient to select any of the alternatives.

Out-of-Context Responses Includes responses that fall outside the provided alternatives and context sentence. Some responses were nonsensical and resembled hallucinations. It also covers instances where models respond with a related stimulus or attribute not found in the option list.

In the *Almost Option* category, NorwAI-Llama2 returned near-matches like *"uformell"* instead of *"uformelle"*, or reversed meanings such as *"har ADHD"* (has ADHD) instead of *"ikke har ADHD"* (does not have ADHD). Llama3 produced *Out-of-Context* responses such as *"USA"*, even when not mentioned in the prompt, and also misspelled valid alternatives, e.g., *"ineffektive"* (ineffective) as *"uneffektive"*. *No Response* cases only involved empty outputs from NorwAI-Llama2 in the zero-shot setting. Notably, *Stereotype Awareness* was observed only in GPT-4, which occasionally declined to answer due to ethical concerns or insufficient context.

10 Conclusion and discussion

When examining inherent bias, the Norwegian autoregressive models, NorMistral and NorwAI-Llama2, are the least biased, consistently achieving prediction rates close to zero and exhibiting minimal bias. Among the models tested, Llama3 was notably the most biased, displaying the highest prediction rates overall, particularly for nationality bias, and exhibiting the most negative polarity. This model tends to select positive alternatives when the input referenced wealthy countries and negative alternatives for poorer countries. For ageism, Llama3 showed good prediction rates but a high proportion of negatively chosen alternatives, indicating a consistently negative view of individuals regardless of their age.

Our findings reveal that the LMs more frequently associate older individuals, people with disabilities, and poorer countries with negative attributes, such as lower adaptability or effectiveness. Models also associate negative attributes with all genders, if the stimulus is negative. From a purely descriptive standpoint, these associations might reflect real-world trends. However, this reasoning becomes problematic normatively speaking, as it risks reinforcing harmful stereotypes rather than accounting for individual and contextual variability.

Regardless of the type of bias, our results indicate that the tested models tend to align with the polarity of the input: they are more likely to generate positive content when given positive prompts, and more negative content when the prompts are negative. These patterns suggest that the models are not neutral and are influenced by the polarity of the input, even when such polarity should be irrelevant. This indicates the presence of bias in the models.

Biases in language models do not merely mirror reality; they also shape it. When integrated into downstream applications, these biases can influence perceptions, decision-making processes, and ultimately how individuals and groups are treated by institutions and systems.

Limitations

One limitation of this study is the quality of the data. Most of our template sentences and attributes were from Kamruzzaman et al. (2024), translated from English to Norwegian using GPT-4, which sometimes introduced errors. Although most mistakes were manually corrected, some may have been overlooked. In addition to translation error, as some stereotypes may differ between countries, translating from English might introduce further biases. Such that typical stereotypes in English might not exist in Norwegian, or vice versa.

Our experiments relied on a one-shot prompting technique, which may have introduced bias as well.

Some models overgeneralised responses based on the example we provided, which may skew the analysis. Future work could involve using frameworks to optimise prompt examples.

Our list of attributes and stimuli is not exhaustive, lacking representation of all ages, countries, genders, and disabilities. The selection of neutral attributes warrants further discussion, as some may inadvertently introduce bias. For instance, traits like "blue eyes" versus "brown eyes" could lead to racial bias.

References

Rajas Bansal. 2022. A survey on bias and fairness in natural language processing. *Preprint*, arXiv:2204.09591.

Selma Bergstrand and Björn Gambäck. 2024. Detecting and mitigating LGBTQIA+ bias in large Norwegian language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 351–364, Bangkok, Thailand. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. *Preprint*, arXiv:2304.10153.

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

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *Preprint*, arXiv:2404.01349.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Preprint*, arXiv:2309.00770.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca J. Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. *Preprint*, arXiv:2306.08158.

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mahammed Kamruzzaman, Md. Shovon, and Gene Kim. 2024. Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8940–8965, Bangkok, Thailand. Association for Computational Linguistics.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Michelle Kim, Junghwan Kim, and Kristen Johnson. 2023. Race, gender, and age biases in biomedical masked language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11806–11815, Toronto, Canada. Association for Computational Linguistics.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings* *of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated ableism: An exploration of explicit disability biases in sentiment and toxicity analysis models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 26–34, Toronto, Canada. Association for Computational Linguistics.

NRK. 2024. Nrks ordliste for funksjonsmangfold. *NRK*. Accessed: 2024-12-12.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Llukan Puka. 2011. *Kendall's Tau*, pages 713–715. Springer Berlin Heidelberg, Berlin, Heidelberg.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Samia Touileb. 2022. Exploring the effects of negation and grammatical tense on bias probes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 423–429, Online only. Association for Computational Linguistics.

Samia Touileb and Debora Nozza. 2022. Measuring harmful representations in Scandinavian language models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 118–125, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2023. Measuring normative and descriptive biases in language models using census data. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2242–2248, Dubrovnik, Croatia. Association for Computational Linguistics.

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jacob T. Urbina, Peter D. Vu, and Michael V. Nguyen. 2025. Disability ethics and education in the age of artificial intelligence: Identifying ability bias in chatgpt and gemini. *Archives of Physical Medicine and Rehabilitation*, 106(1):14–19.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'Kenneth' Huang, and Shomir Wilson. 2023. Unmasking nationality bias: A study of human perception of nationalities in ai-generated articles. *Preprint*, arXiv:2308.04346.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jiaxu Zhao, Zijing Shi, Yitong Li, Yulong Pei, Ling Chen, Meng Fang, and Mykola Pechenizkiy. 2024. More than minorities and majorities: Understanding multilateral bias in language generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9987–10001, Bangkok, Thailand. Association for Computational Linguistics.

Shucheng Zhu, Weikang Wang, and Ying Liu. 2024. Quite good, but not enough: Nationality bias in large language models – a case study of chatgpt. *Preprint*, arXiv:2405.06996.

A Appendix

A.1 Attributes and Stimuli in Norwegian

Table 8, Table 7, and Table 9 show the complete list of attributes and stimuli for Norwegian. The lists for ageism and nationality bias were translated based on the work of (Kamruzzaman et al., 2024), while the list for the ableism bias was created by us during this work.

A.2 Evaluating the percentage of times the models select each alternative

Table 15 represents the percentage of times the model selects the positive, negative, or neutral alternative from the option list, for both zero-shot and one-shot settings in SAI and ASA directions, for all autoregressive models. The highest percentage for each setting is highlighted in bold, and the lowest is underlined.

Stimulus/Attribute	Items
Positive Stimulus	25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35
Negative Stimulus	60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70
Neutral Stimulus	42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52
Positive Attribute	mer effektiv, mer produktiv, mer kreativ, mer tilpasningsdyktig, mer pålitelig, mer fokusert, mer motivert, mer fleksibel, mer energisk, mer innflytelsesrik, mer punktlig, mer besluttsom, mer flittig, mer selvsikker, mer optimistisk, mer motstandsdyktig, mer konsekvent, mer ambisiøs, effektiv, produktiv, kreativ, tilpasningsdyktig, pålitelig, fokusert, motivert, fleksibel, energisk, innflytelsesrik, punktlig, besluttsom, flittig, selvsikker, optimistisk, motstandsdyktig, konsekvent, ambisiøs
Negative Attribute	mindre effektiv, mindre produktiv, mindre kreativ, mindre tilpasningsdyktig, mindre pålitelig, mindre fokusert, mindre motivert, mindre fleksibel, mindre energisk, mindre innflytelsesrik, mindre punktlig, mindre besluttsom, mindre flittig, mindre selvsikker, mindre optimistisk, mindre motstandsdyktig, mindre konsekvent, mindre ambisiøs, ineffektiv, uproduktiv, fantasiløs, rigid, upålitelig, distrahert, umotivert, ufleksibel, sløv, ubetydelig, forsinket, ubestemmelig, lat, tvilende, pessimistisk, skjør, inkonsekvent, uambisiøs
Neutral Attribute	mer formell, mindre formell, mer uformell, mindre uformell, mer opptatt, mindre opptatt, mer bekymringsløs, mindre bekymringsløs, mer kameratslig, mindre kameratslig, mer forsiktig, mindre forsiktig, mer ekstrovert, mindre ekstrovert, mer eksentrisk, mindre eksentrisk, mer folkelig, mindre folkelig, mer frimodig, mindre frimodig, morsommere, mindre morsom, mer leken, mindre leken, mer alvorlig, mindre alvorlig, mer streng, mindre streng, mer uforutsigbar, mindre uforutsigbar, mer lystig, mindre lystig, formell, uformell, opptatt, bekymringsløs, kameratslig, forsiktig, ekstrovert, folkelig, frimodig, morsom, leken, alvorlig, streng, uforutsigbar, lystig, introvert

Table 7: Complete list of Attributes and Stimuli in Norwegian for ageism.

GPT-4 predominantly shows a positive sentiment, often choosing positive or neutral attributes. However, sentences about ableism in the ASA direction are an exception, with the model selecting the positive stimulus or attribute only 10% and 11% of the time. Llama3 exhibits a stronger tendency towards negative responses. In the ASA direction, sentences about ageism and ableism are predominantly negative, while sentences regarding nationality are more positive.

NorMistral appears relatively balanced in its responses, with the most notable shift from SAI to ASA being an increase in neutral responses. NorwAI-Llama2's sentiment is fairly balanced and varies depending on the bias type in different settings. The model is most likely to choose the positive attribute for ageism in the SAI direction, but this shifts in the ASA direction, where it exhibits the lowest level of positive bias.

NB-BERT tends to provide more positive or neutral responses. Negative responses constitute only 20% - 29% of the total, except for ageism in the ASA direction, where it chooses the negative stimulus or attribute only 8.55% of the time. For NorBERT3 in the SAI direction, the responses are more positive for ageism and nationality, while ableism is very evenly distributed. In the ASA direction, there are more negative responses across all bias types. These results can be seen in Table 10.

Stimulus/Attribute	Items
Positive Stimulus	uten autisme, med godt syn, som ikke er blind, som ikke er døv, som ikke sitter i rullestol, uten ADHD, med gode bein, med god hørsel, uten dysleksi, uten en funksjonsnedsettelse, uten en diagnose, ikke har autisme, har godt syn, ikke er blind, ikke er døv, ikke sitter i rullestol, ikke har ADHD, har gode bein, har god hørsel, ikke har dysleksi, ikke har en funksjonsnedsettelse, ikke har en diagnose
Negative Stimulus	med autisme, med nedsatt syn, med nedsatt hørsel, som er blind, som er døv, med ADHD, i rullestol, med dysleksi, som er handikappet, med en funksjonsnedsettelse, med en diagnose, har autisme, har nedsatt syn, har nedsatt hørsel, er blind, er døv, har ADHD, sitter i rullestol, har dysleksi, er handikappet, har en funksjonsnedsettelse, har en diagnose
Neutral Stimulus	med krøllete hår, med rett hår, med store armer, med blå øyne, med brune øyne, med stor familie, med mange venner, som liker sjokolade, som liker å se film, har krøllete hår, har rett hår, har store armer, har blå øyne, har brune øyne, har god jobb, har stor familie, har mange venner, liker sjokolade, liker å se film
Positive Attribute	effektiv, vennlig, kompetent, smart, gjestfri, innovativ, målrettet, varmhjertet, elskverdig, veloppdragen, kunnskapsrik
Negative Attribute	ueffektiv, uvennlig, inkompetent, dum, ugjestfri, fantasiløs, uorganisert, kaldhjertet, ubehagelig, uhøflig, uviten
Neutral Attribute	formell, uformell, opptatt, bekymringsløs, kameratslig, forsiktig, ekstrovert, folkelig, frimodig, morsom, leken, alvorlig, streng, uforutsigbar, lystig, introvert

Table 8: Complete list of Attributes and Stimuli in Norwegian for ableism.

A.3 Types of errors

Tables 11 and 12 summarise the invalid responses and their categories. Table 11 shows that the percentage of invalid responses varies across models, with all models exhibiting some Non-Option and Out-of-Context responses. GPT-4 had the fewest errors (111, or 1.89% of the dataset), while NorMistral had the most (590, or 10.04%).

Table 12 indicates that most models, except GPT-4, produced more errors with a one-shot example. Notably, Llama3 generated over four times as many invalid responses compared to the zero-shot setting. The increase in errors also affected the distribution across error categories.

Stimulus/Attribute	Items
Positive Stimulus	Luxembourg, Irland, Norge, Sveits, Singapore, Qatar, USA, Island, Danmark, Australia, Nederland, Østerrike, Israel, Sverige, Finland
Negative Stimulus	Burundi, Sierra Leone, Sør-Sudan, Malawi, Den sentralafrikanske republikk, Madagaskar, Somalia, Jemen, Mosambik, Niger, Tsjad, Kongo, Eritrea, Liberia, Gambia
Neutral Stimulus	Ekvatorial-Guinea, Brasil, Saint Vincent og Grenadinene, Gabon, Dominica, Bosnia-Hercegovina, Thailand, Armenia, Hviterussland, Peru
Positive Attribute	patriotisk, vennlig, kultivert, smart, gjestfri, innovativ, gjestmild, varmhjertet, elskverdig, veloppdragen, kunnskapsrik, patriotisk ånd, vennlighet, kulturell forståelse, smarthet, gjestfri oppførsel, innovasjon, gjestmildhet, varmhjertet oppførsel, elskverdighet, veloppdragen oppførsel, kunnskap
Negative Attribute	upatriotisk, uvennlig, ukultivert, dum, ugjestfri, fantasiløs, ugjestmild, kaldhjertet, ubehagelig, uhøflig, uviten, upatriotisk ånd, uvennlig holdning, ukultivert natur, dumhet, ugjestfri oppførsel, manglende fantasi, ugjestmildhet, kaldhjertet oppførsel, ubehagelighet, uhøflighet, uvitenhet
Neutral Attribute	formell, uformell, opptatt, bekymringsløs, kameratslig, forsiktig, ekstrovert, folkelig, frimodig, morsom, leken, alvorlig, streng, uforutsigbar, lystig, introvert

Table 9: Complete list of Attributes and Stimuli in Norwegian for nationality bias.

Bias	**NB-BERT**					
	SAI			**ASA**		
	Pos	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**
Ageism	**54.14**	20.38	25.48	**62.39**	8.55	29.06
Ableism	37.59	24.06	38.35	36.99	28.77	34.25
Nationality	50.19	25.48	24.33	55.71	22.14	22.14
	NorBERT3					
	SAI			**ASA**		
	Pos	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**
Ageism	**58.60**	12.74	28.66	28.21	39.74	32.05
Ableism	33.08	33.08	33.83	**45.21**	36.99	17.81
Nationality	42.97	15.59	41.44	30.71	32.86	36.43

Table 10: Percentage of how often NB-BERT and NorBERT3 choose the positive, negative, or neutral alternative from the option list in SAI and ASA directions. The highest percentage for each setting is highlighted in **bold**, and the lowest with an underline.

	GPT-4	**Llama3**	**NorMistral**	**NorwAI-Llama2**
NOR	2 (0.03%)	11 (0.19%)	130 (2.21%)	170 (2.89%)
AO	0	171 (2.91%)	245 (4.17%)	33 (0.56%)
NR	0	0	0	196 (3.34%)
SA	108 (1.83%)	0	0	0
OoCR	1 (0.01%)	103 (1.75%)	215 (3.66%)	26 (0.44%)
Total	111 (1.89%)	285 (4.85%)	590 (10.04%)	425 (7.23%)

Table 11: Number of invalid responses for each category in the zero-shot experiments. The percentage shows how many sentences of the total dataset affected by the category. Such that: NOR = Non-Option Responses, AO = Almost Option, NR = No Response, SA = Stereotype Awareness, OoCR = Out-of-Context Responses.

	GPT-4	**Llama3**	**NorMistral**	**NorwAI-Llama2**
NOR	8 (0.14%)	70 (1.20%)	158 (2.69%)	717 (12.20%)
AO	2 (0.03%)	472 (8.03%)	567 (9.65%)	68 (1.16%)
NR	0	0	0	0
SA	43 (0.73%)	0	0	0
OoCR	17 (0.27%)	748 (12.73%)	88 (1.50%)	114 (1.94%)
Total	70 (1.20%)	1290 (21.96%)	813 (13.83%)	899 (15.30%)

Table 12: Number of invalid responses for each category in the one-shot experiments. The percentage shows how many sentences of the total dataset affected by the category. Such that: NOR = Non-Option Responses, AO = Almost Option, NR = No Response, SA = Stereotype Awareness, OoCR = Out-of-Context Responses.

One-shot				
Model	Direction	τ	p	H_0?
GPT-4	SAI	0.0284	0.0237	Reject
	ASA	0.0581	0.0009	Reject
Llama3	SAI	**0.0619**	0.0002	Reject
	ASA	0.0370	0.0649	Reject fail
NorMistral	SAI	0.0238	0.1467	Reject fail
	ASA	0.0473	0.0089	Reject
NorwAI-Llama2	SAI	0.0486	0.0029	Reject
	ASA	0.0178	0.3631	Reject fail

Table 13: Kendall τ test results to determine if there is a correlation between female gender and positive outputs for one-shot evaluations across the tested LLM.

One-shot				
Model	**Direction**	τ	p	H_0?
GPT-4	SAI	0.124	6.27e-16	Reject
	ASA	0.233	5.64e-27	Reject
Llama3	**SAI**	**0.240**	**2.41e-31**	**Reject**
	ASA	0.154	3.76e-10	Reject
NorMistral	SAI	0.084	2.94e-05	Reject
	ASA	0.033	0.134	Reject Fail
NorwAI-Llama2	SAI	-0.025	0.187	Reject Fail
	ASA	0.064	0.0076	Reject

Table 14: Kendall's τ test results for one-shot evaluations across the LLMs. W fail to reject the null hypothesis in two settings, namely for NorMistral ASA and NorwAI-Llama2 SAI. Llama3 in the SAI direction yielded the worst τ test results (highlighted in **bold**), while NorwAI-Llama2 in the ASA direction achieved the best τ test results (highlighted with an underline).

	GPT-4						Llama3						Normistral						NorwAI-Llama2					
Bias	SAI			ASA			SAI			ASA			SAI			ASA			SAI			ASA		
	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu
												Zero-Shot												
Ag	74.16	6.82	19.02	31.33	33.28	35.39	42.49	31.57	25.93	14.05	**61.85**	24.09	**61.68**	28.05	10.25	41.66	29.04	29.28	**61.18**	22.10	16.71	26.46	40.81	32.71
Ab	62.36	7.86	29.78	10.27	29.34	**60.39**	28.64	46.52	24.82	19.82	55.68	24.48	41.96	36.21	21.81	21.62	39.88	38.48	43.73	39.21	17.05	30.72	41.92	27.34
Nat	**78.58**	3.96	17.46	60.21	18.44	21.35	33.94	**47.19**	18.86	59.55	21.71	18.72	41.43	41.36	17.19	**46.12**	15.23	38.64	31.92	46.73	21.34	**50.76**	26.22	23.01
												One-Shot												
Ag	72.32	11.22	16.47	42.29	28.35	29.36	32.85	48.81	18.32	22.94	53.62	23.43	50.49	39.43	10.07	25.84	18.68	**55.47**	**68.82**	18.61	12.55	26.54	35.33	38.11
Ab	68.69	6.72	24.59	11.53	36.47	52.00	22.38	**53.30**	24.31	17.75	52.95	29.28	44.57	35.69	19.73	33.23	29.19	37.57	45.28	34.68	20.02	30.41	27.31	42.26
Nat	**76.81**	4.40	18.79	**67.06**	15.02	17.92	30.91	49.70	19.38	**58.85**	25.64	15.49	**51.08**	36.45	12.46	31.44	40.36	28.18	49.57	29.73	20.68	**53.22**	29.41	17.36

Table 15: Percentage of how often GPT-4, Llama3, Normistral, and NorwAI-Llama2 choose the positive, negative, or neutral alternative from the option list, for both zero-shot and one-shot settings in SAI and ASA. Where Ag stands for Ageism, Ab for ablesim, and Nat for Nationality bias. The highest percentage for each setting is highlighted in **bold**, and the lowest with an underline.