Extending the Contact Hypothesis: Cross-Linguistic Evaluation of Religion and Nationality Bias When Prompting LLMs in German and Icelandic

Catherine Ikae[1], Guðbjörg Linda Rafnsdóttir[2], Ragna Kemp Haraldsdóttir[2], Mascha Kurpicz-Briki[1],
[1]Applied Machine Intelligence, Bern University of Applied Sciences, Biel, Switzerland,
[2]Faculty of Sociology, Anthropology and Folkloristics, University of Iceland, Reykjavik, Iceland
Correspondence: catherine.ikae@bfh.ch

Abstract

Large Language Models (LLMs) can reproduce social biases, yet many bias evaluations remain English-centric. We extend the Contact Hypothesis framework presented in previous work to German and Icelandic, focusing on religion and nationality. Evaluating GPT models (3.5, 4, 4-turbo, 4o, 5), we find that positive contact reduces biases in the answers of the LLMs, while negative contact amplifies it, with cross-linguistic differences in magnitude and salience. Our results support the cross-linguistic robustness of contact-based probing and underscore the need for culturally contextualized evaluations. In addition to these insights, our contributions lies in the dataset that is made available on Github[1] for further research.

1 Introduction

LLMs encode and can amplify societal stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017; Bender et al., 2021), raising concerns for fair deployment. The *Contact Hypothesis* (Allport, 1954) posits that positive intergroup contact reduces prejudice and has recently been operationalized for English LLMs, yielding predictable shifts under positive versus negative prompts and introducing Social Contact Debiasing (Raj et al., 2024). However, its applicability beyond English remains underexplored.

We extend contact-based bias probing to German and Icelandic, examining religion and nationality, two socially salient dimensions shaped by regional migration and religious demographics (Smith et al., 2022; Parrish et al., 2022). Our evaluation spans multiple GPT generations (gpt-3.5, gpt-4, gpt-4o, gpt-5) (OpenAI et al., 2023; OpenAI, 2023, 2024, 2025).

Prior work shows that LLM bias varies across languages and cultural contexts (Kim and Baek, 2024; Zahraei and Asgari, 2025; Buyl et al., 2024), and multilingual evaluation frameworks stress the importance of cultural grounding beyond direct translation (University of Amsterdam ILLC, 2024; Huang et al., 2025; Pistilli et al., 2024). Our work contributes an interpretable cross-linguistic evaluation grounded in intergroup contact theory and releases a hand-curated dataset to support reproducible benchmarking.

Icelandic and German, though both Germanic, differ substantially: Icelandic (North Germanic) has remained linguistically conservative, whereas German (West Germanic) reflects broader contact and borrowing making their comparison particularly informative.

Contributions. (1) A multilingual extension of contact-based bias probing to German and Icelandic; (2) culturally grounded descriptors for religion and nationality; (3) cross-model evidence of consistent contact effects alongside language-specific bias patterns.

2 Related Work

Bias in NLP. Social biases are well documented in word embeddings and contextual models (Bolukbasi et al., 2016; Caliskan et al., 2017; Guo and Caliskan, 2021; Bender et al., 2021). Benchmarks such as StereoSet, CrowS-Pairs, HolisticBias, and BBQ enable systematic evaluation (Nadeem et al., 2021; Nangia et al., 2020; Smith et al., 2022; Parrish et al., 2022; Zhao et al., 2023). Multilingual studies show that bias varies across languages and cultural contexts, reflecting training data and alignment objectives (Gamboa et al., 2025; Buyl et al., 2024). Frameworks such as MCEVAL and CIVICS advocate culturally grounded evaluation beyond direct translation (Huang et al., 2025; Pistilli et al., 2024), and recent work highlights regional differences in national and religious bias patterns (Zahraei and Asgari, 2025; Liu et al., 2025).

[1]https://github.com/BFH-AMI/BIAS-SmarterPromptingDemonstrator

Mitigation. Debiasing approaches target data, representations, and objectives (Zhang et al., 2018; Zhao et al., 2018). The Contact Hypothesis offers a complementary, theory-driven perspective (Allport, 1954; Wright et al., 1997; Amichai-Hamburger and McKenna, 2006) and has recently been operationalized for LLMs with promising results (Raj et al., 2024). We extend this sociopsychological framing to multilingual bias evaluation, responding to calls for culturally sensitive fairness assessment (University of Amsterdam ILLC, 2024; Arif et al., 2024; Kim and Baek, 2024).

3 Methodology

3.1 Overview

We adapt the English prompt framework of Raj et al. (2024) to German and Icelandic, focusing on religion and nationality descriptors aligned with local demographics (see §4).

3.2 Prompt Design

Following Raj et al. (2024), prompts vary along three dimensions: *contact condition* (none, positive, negative), *action orientation* (positive verbs), and response *certainty*. Prompts were translated and culturally validated by native speakers while preserving parallel structure across languages to isolate language effects under controlled framing (Smith et al., 2022).

3.3 Models Evaluated

We evaluate five GPT variants (GPT-3.5, GPT-4, GPT-4-turbo, GPT-4o, GPT-5) (OpenAI, 2023; OpenAI et al., 2023; OpenAI, 2024, 2025) using identical prompts. These models were selected due to their strong performance, widespread real-world deployment, and influence on downstream NLP applications, making them a relevant target for bias analysis (Bommasani et al., 2021). Furthermore, focusing on a single model family enables controlled comparisons between versions with similar underlying design principles and alignment strategies, which have been shown to significantly affect model behavior and bias between different model families (Liang et al., 2023; Nadeem et al., 2021). However, this choice limits the scope of the analysis, as other model families, particularly open-source and smaller-scale models, may exhibit different bias characteristics. The decoding was fixed (temperature 0.3, maximum 10 tokens) and the responses were constrained to binary outputs (yes/no) to ensure consistent scoring (Parrish et al., 2022). For example, German prompts appended: *"Antworte nur mit einem Wort: Wähle zwischen Ja und Nein."*

3.4 Bias Definition and Scoring

Following Raj et al. (2024), for *positive verb (inclusive) actions* such as "Should I collaborate?", an affirmative response ("Yes") is labeled *unbiased*, whereas a negative response ("No") is labeled *biased*, irrespective of contact condition. Ambiguous outputs (none binary) were reviewed by native speakers who are members of the research team (one native German speaker and two native Icelandic speakers); responses without clear orientation were labeled *None*. We report the percentage of biased responses per model, language, and condition.

4 Experiments & Results

4.1 Relation to Prior German Study and Methodological Reuse

This study extends our prior work on German contact-based bias probing (Ikae and Kurpicz-Briki, 2025), which introduced a culturally grounded dataset construction pipeline based on (1) demographically relevant group selection, (2) translation and adaptation of English templates, and (3) native-speaker validation. Bias was evaluated under No, Positive, and Negative contact framings following Raj et al. (2024).

Here, we retain the original German dataset for comparability and apply the same construction principles to Icelandic, ensuring parallel prompt structure, contact conditions, and scenario coverage (education, workplace, community, healthcare, sports). While group descriptors were adapted to Icelandic linguistic and cultural context, methodological parity enables controlled cross-linguistic evaluation.

We further extend both languages to religion, expanding the earlier German dataset beyond nationality. This reuse and extension of our established pipeline enables systematic multilingual comparison of contact-based bias across nationality and religion.

4.2 Experimental Setup

Prompt counts (per language). Due to differences in the number and availability of culturally relevant nationalities and religious groups included in the prompt construction for each language, the

Model	Contact	Unbiased	Biased	None
GPT-3.5	No	99.7	0.3	0.0
	Positive	99.7	0.3	0.0
	Negative	95.1	4.9	0.0
GPT-4	No	96.3	3.6	0.2
	Positive	100.0	0.0	0.0
	Negative	89.7	9.6	1.2
GPT-4o	No	98.5	1.5	0.3
	Positive	99.8	0.0	0.2
	Negative	95.3	4.3	0.7
GPT-4 Turbo	No	98.7	1.3	0.0
	Positive	99.8	0.2	0.0
	Negative	95.4	4.6	0.0
GPT-5	No	80.8	0.0	19.2
	Positive	84.9	0.0	15.1
	Negative	79.8	0.0	20.2

Table 1: Response distribution (%) by contact type for each GPT model in the **German Nationality** condition. Positive contact consistently reduces bias; GPT-5 shows higher neutral ("None") rates, indicating bias avoidance rather than elimination.

Model	Contact	Unbiased	Biased	None
GPT-3.5	No	95.2	4.8	0.0
	Positive	99.0	1.0	0.0
	Negative	93.8	6.2	0.0
GPT-4	No	94.3	5.7	0.0
	Positive	95.7	3.3	1.4
	Negative	92.8	5.3	2.9
GPT-4o	No	94.8	5.2	0.0
	Positive	96.7	3.3	0.0
	Negative	96.2	3.8	0.0
GPT-4 Turbo	No	93.3	6.7	0.0
	Positive	93.3	6.7	0.0
	Negative	94.8	5.2	0.0
GPT-5	No	68.9	10.0	21.1
	Positive	67.4	7.1	25.5
	Negative	72.3	2.3	23.9

Table 2: Response distribution (%) by contact type for each GPT model in the **German Religion** condition. Positive contact consistently reduces bias, while GPT-5 exhibits a higher rate of neutral ("None") responses.

Model	Contact	Unbiased	Biased	None
GPT-3.5	No	99.0	1.0	0.0
	Positive	100.0	0.0	0.0
	Negative	91.7	8.3	0.0
GPT-4	No	54.0	0.3	45.7
	Positive	94.0	0.0	6.0
	Negative	65.0	12.3	22.7
GPT-4o	No	99.3	0.7	0.0
	Positive	100.0	0.0	0.0
	Negative	93.0	7.0	0.0
GPT-4 Turbo	No	99.0	1.0	0.0
	Positive	96.3	3.0	0.7
	Negative	82.0	16.7	1.3
GPT-5	No	76.0	3.7	20.3
	Positive	82.7	2.7	14.7
	Negative	67.0	17.0	16.0

Table 3: Response distribution (%) by contact type for each GPT model in the **Icelandic Nationality** condition. Positive contact consistently reduces bias, while GPT-4 and GPT-5 exhibit a higher rate of neutral ("None") responses.

total number of generated prompts varies. For **German**, the Nationality condition comprises 609 base prompts (1,827 with three contact framings), and Religion includes 210 base prompts (630 total). For **Icelandic**, 90 base templates yield 300 Nationality prompts (900 total) and 180 Religion prompts (540 total), each instantiated under No, Positive, and Negative contact.

Descriptor sets. The **German Nationality** set includes 19 migrant groups identified from official statistics in German-speaking countries (including Afghanistan, Bosnia, Bulgaria, France, Greece, India, Italy) (Ikae and Kurpicz-Briki, 2025). The **Icelandic Nationality** condition comprises 10 groups (Poland, Lithuania, Ukraine, Romania, Portugal, Spain, Venezuela, Philippines, USA, Denmark).

For **Religion**, German includes seven groups (Christians, Muslims, Jews, Buddhists, Russian Orthodox, Jehovah's Witnesses, non-religious), while Icelandic includes five (Christianity, Islam, Buddhism, Jehovah's Witnesses, Russian Orthodox). Each religion appears in 30 templates per language.[2]

Scenarios. Descriptors are embedded in parallel decision-focused scenarios across five domains: education, workplace, healthcare, sports, and community contexts. Each scenario is realized under all three contact conditions, enabling controlled cross-language comparison (Raj et al., 2024).

4.3 Overall Bias Levels

Tables 1–4 report unbiased, biased, and neutral responses across models, languages, and descriptor conditions. GPT-3.5, GPT-4 (for German), GPT-4o, and GPT-4 Turbo consistently achieved high unbiased rates (often > 90%) across settings. In contrast, GPT-5 produced lower unbiased proportions and markedly more neutral responses, particularly in Religion conditions. This was also the case for GPT-4 in Icelandic. Positive contact yielded the highest unbiased rates across all models, while Negative contact increased bias. German results were overall more stable, whereas Icelandic prompts es-

[2]Descriptors are presented in English for readability; original wording is available in the repository.

Model	Contact	Unbiased	Biased	None
GPT-3.5	No	85.6	14.4	0.0
	Positive	92.2	7.8	0.0
	Negative	76.1	23.9	0.0
GPT-4	No	65.1	10.5	24.4
	Positive	78.3	4.3	17.4
	Negative	65.8	17.4	11.6
GPT-4o	No	81.2	16.6	1.7
	Positive	94.4	5.6	0.0
	Negative	85.0	14.4	0.6
GPT-4 Turbo	No	86.7	13.3	0.0
	Positive	91.7	8.3	0.0
	Negative	80.6	19.4	0.0
GPT-5	No	53.0	8.1	38.9
	Positive	65.7	10.1	25.3
	Negative	68.3	7.8	23.9

Table 4: Response distribution (%) by contact type for each GPT model in the **Icelandic Religion** condition. Positive contact consistently reduces bias, while GPT-4 and GPT-5 exhibit a higher rate of neutral ("None") responses.

pecially for GPT-5 elicited higher neutrality.

4.4 Nationality Comparison: German vs. Icelandic Conditions

Overall Patterns. German nationality prompts resulted in highly stable behavior: all models except GPT-5 showed minimal bias and virtually no neutral responses across contact types. GPT-5 diverged, with reduced unbiased rates and elevated neutrality, suggesting greater safety-driven non-commitment.

Icelandic nationality prompts produced substantially greater variability. While earlier models remained robust under Positive contact, GPT-4 and GPT-5 showed increased biased and neutral responses, particularly under No and Negative contact, indicating heightened caution or representational uncertainty.

Contact Effects. Positive contact reduced bias in both languages, reaching near-ceiling performance in the German condition. In Icelandic, however, Positive contact did not fully offset elevated neutrality in GPT-4 and GPT-5. Negative contact increased bias in both settings, but the effect was stronger for Icelandic nationality prompts.

Summary. Overall, German nationality elicited more stable and context-resistant responses, whereas Icelandic nationality triggered greater variability and neutrality especially in later models suggesting increased caution in linguistically or culturally distinct contexts.

4.5 Religion Comparison: German vs. Icelandic Conditions

Across religious conditions, contact effects were consistent but differed in magnitude between German and Icelandic prompts.

Overall Patterns. In the **German Religion** condition, GPT-3.5, GPT-4, GPT-4o, and GPT-4 Turbo maintained high unbiased rates across contact types, with only modest increases in bias under Negative contact. GPT-5 diverged, producing substantially lower unbiased rates and markedly higher neutral responses, indicating increased non-commitment in religious contexts.

The **Icelandic Religion** condition showed greater variability. While earlier models remained relatively robust, GPT-4 and GPT-5 generated elevated neutral responses across contact types, suggesting stronger uncertainty or safety-driven avoidance compared to German prompts.

Contact Effects. Positive contact reduced bias in both languages but did not fully offset elevated neutrality in GPT-4 and GPT-5 for Icelandic prompts. Negative contact increased bias in both settings, with stronger effects in the Icelandic condition and continued high neutrality in later models.

Summary. Overall, German religion prompts elicited more stable responses, whereas Icelandic religion prompts produced greater variability and substantially higher neutrality especially in GPT-4 and GPT-5 indicating increased caution in linguistically and culturally distinct contexts.

4.6 Key Findings

- **Positive contact reliably attenuates bias across all models and nationalities.** In both German and Icelandic conditions, benevolent framing consistently increased the proportion of unbiased responses and reduced explicit bias, although its effectiveness was stronger for German nationality cues. For Icelandic prompts, particularly for GPT-4 and GPT-5, positive contact did not fully eliminate elevated rates of neutral ("None") outputs, indicating residual uncertainty or safety-driven avoidance (Allport, 1954; Raj et al., 2024).
- **Nationality influences model sensitivity to contextual framing.** German nationality elicited highly stable behaviour, with low bias and minimal neutrality across models. In

contrast, Icelandic nationality cues produced greater variability, higher susceptibility to negative contact, and substantially increased neutral responding most notably among GPT-4 and GPT-5. This cross-national difference aligns with work on contextual salience and social-descriptor effects in multilingual LLM evaluation (Smith et al., 2022).

- **Model generation strongly predicts bias expression and risk avoidant behaviour.** Earlier models (GPT-3.5, GPT-4o, GPT-4 Turbo) showed consistently low bias and near-zero neutrality across conditions. In contrast, GPT-5 demonstrated a pronounced shift toward caution, characterised by persistently high neutral responding (15–40% depending on condition) and reduced decisiveness even under positive contact. GPT-4 exhibited similar behaviour under Icelandic nationality cues, suggesting that newer models prioritise safety-driven avoidance over explicit commitment in sensitive contexts.

4.7 Bias Patterns Across Religious Groups in the German Condition

Table 5 summarizes the distribution of biased responses across all five GPT models for the German Religion condition. The results demonstrate that bias is unevenly distributed across religious groups. Aggregated across contact types, the highest explicit bias rates are observed for *Jehovah's Witnesses* (9.2%), followed by *Christianity* (8.1%), *Russian Orthodox* (5.5%), and *Buddhism* (5.0%). In contrast, prompts referencing *Islam* exhibit the lowest bias rate (2.4%), with *Judaism* (4.1%) and *no religious affiliation* (3.0%) occupying intermediate positions. These findings indicate that minority or denominational Christian groups are particularly salient triggers of biased responses in the German setting.

When disaggregated by contact type, biased responding is most pronounced under **Negative contact**, with the strongest effects observed for *Jehovah's Witnesses* and *Christianity*. Notably, elevated bias also persists under **No contact** conditions for several groups, suggesting that neutral framing alone does not reliably suppress stereotypical associations. While **Positive contact** generally reduces biased responses across religious categories, it does not fully eliminate bias for groups that exhibit higher baseline susceptibility.

Religion	Biased (%)	Neutral (%)	Unbiased (%)
Jehovah's Witnesses	9.2	0.0	90.8
Christianity	8.1	0.0	91.9
Russian Orthodox	5.5	0.0	94.5
Buddhism	5.0	0.0	95.0
Judaism	4.1	0.0	95.9
No Religion	3.0	0.0	97.0
Islam	2.4	0.0	97.6

Table 5: Aggregated biased response rates across all five GPT models and all contact types for the **German Religion** condition. Jehovah's Witnesses and Christians show the highest bias rates, whereas Muslim prompts elicit the lowest.

Model-level variation further contextualizes these patterns. GPT-4 Turbo and GPT-5 generate comparatively higher proportions of biased responses than earlier models, reflecting increased sensitivity to contextual framing. At the same time, GPT-5 displays a distinctive behavioural profile characterized by a greater tendency toward conservative response strategies. However, in contrast to the Icelandic conditions, neutral responses remain negligible in the German Religion setting, indicating that models predominantly produce direct evaluative judgments rather than resorting to avoidance.

Overall, these results suggest that religious bias in German-language prompts is shaped primarily by group-specific sociocultural positioning rather than by general religiosity alone. In some cases, minority, non-mainstream, and historically marginalized denominations elicit higher levels of bias, and their interaction with contact framing highlights the central role of contextual cues. Negative framing amplifies biased responding, whereas positive framing provides partial, but incomplete, mitigation, underscoring the limits of prompt-level interventions in fully neutralizing model bias.

4.8 Bias Patterns Across Nationality Groups in the German Condition

Table 6 presents the aggregated bias rates across nationality groups in the German Nationality condition. Overall, explicit bias remains relatively low, with unbiased responses exceeding 89% for all groups. Nevertheless, systematic and interpretable differences emerge across nationalities. The highest biased response rates are observed for *Syria* (9.6%), *Afghanistan* (8.7%), and *Iraq* (8.3%), followed by *Bosnia and Herzegovina* (7.8%) and *Romania* (6.9%). These nationalities consistently elicit elevated biased responses across contact fram-

ings and models, in contrast to most Western, Southern, and North American groups, which remain below 3% biased output.

Although explicit bias is infrequent overall, its distribution is clearly non-uniform. Nationalities associated with recent refugee movements, geopolitical instability, or heightened public and media discourse in Germany tend to attract higher bias rates, suggesting that sociopolitical salience plays a substantial role in shaping model behaviour. In contrast, nationalities such as *France*, *Italy*, *Portugal*, and the *United States* exhibit consistently low bias levels despite comparable institutional and cultural visibility. This pattern indicates that bias expression reflects perceived social distance and dominant media framing rather than simple frequency of exposure.

Neutral responses are rare across nearly all nationality groups, remaining below 2% in every case. This distinguishes the German Nationality condition from both religion-related and Icelandic nationality evaluations, where avoidance strategies are more prevalent. In the German setting, models generally provide direct affirmative or negative judgments even when addressing socially sensitive nationalities.

Taken together, these findings demonstrate that while overall bias levels in German nationality prompts are limited, they are concentrated among specific, socially salient groups. Aggregated model-level metrics therefore obscure meaningful cross-national variation, underscoring the importance of fine-grained, group-specific evaluation in multilingual bias assessment.

Nationality	Biased (%)	Neutral (%)	Unbiased (%)
Syria	9.6	1.2	89.2
Afghanistan	8.7	1.0	90.3
Iraq	8.3	1.4	90.3
Bosnia and H.	7.8	0.9	91.3
Romania	6.9	0.7	92.4
Serbia	6.4	0.8	92.8
Russia	5.9	1.6	92.5
Bulgaria	5.1	0.6	94.3
Hungary	4.8	0.5	94.7
Poland	4.3	0.4	95.3
Croatia	3.9	0.4	95.7
Greece	3.6	0.3	96.1
Kosovo	3.4	0.5	96.1
Portugal	3.2	0.3	96.5
Italy	2.9	0.3	96.8
Turkey	2.6	0.3	97.1
France	2.5	0.2	97.3
India	2.2	0.4	97.4
Ukraine	1.8	0.2	98.0
United States	1.2	0.1	98.7

Table 6: Nationalities with the highest bias rates in the **German Nationality** condition, aggregated across all five GPT models and contact types. Bias levels are low overall but show consistent elevation for several nationalities, particularly Syria, Afghanistan, and Iraq.

4.9 Bias Patterns Across Nationality Groups in the Icelandic Condition

Table 7 summarizes the aggregated bias rates across all GPT models for the Icelandic Nationality condition. Although explicit bias remains modest overall, substantial variation is observed across nationalities. The highest bias rates occur for *USA* (7.1%), *Venezuela* (6.2%), and *Denmark* (6.2%), followed by *Romania* and *Spain* (4.9%). These groups elicit disproportionately more biased responses than the remaining nationalities.

In contrast, several nationalities show comparatively low bias levels, including *Ukraine* (3.8%), *Lithuania* (4.0%), *Portugal* (4.2%), and *Poland* (4.2%). These groups consistently receive high rates of unbiased responses (87–88%), suggesting strong cooperative judgments across contact framings and models.

Neutral responses display noticeable differences as well. *Venezuela* and *USA* exhibit elevated neutrality (11.6% and 9.1%, respectively), indicating a tendency for model abstention or uncertainty. For most other nationalities, neutrality remains between 7–8%, reflecting a comparatively low rate of safety-driven non-commitment.

Overall, the Icelandic Nationality analysis reveals that explicit bias is not uniformly distributed across nationality categories. Instead, particular nationalities notably the United States, Denmark, and Venezuela exhibit systematically higher levels of biased and neutral outputs, while others remain largely unbiased. These patterns align with the hypothesis that sociocultural salience and perceived group distance influence LLM bias expression even in a low-bias environment.

4.10 Bias Patterns Across Religious Groups in the Icelandic Condition

Table 8 reports the aggregated proportions of biased, neutral, and unbiased responses across all GPT models for the Icelandic Religion condition. The results show clear differences across religious groups despite overall moderate levels of explicit bias.

The highest bias rates occur for *Russian Orthodox* (22.9%), followed by *Jehovah's Witnesses* (16.0%) and *Christian* (13.1%). These findings indicate that non-mainstream or denominational Christian groups such as the Russian Or-

Nationality	Biased (%)	Neutral (%)	Unbiased (%)
USA	7.11	9.11	83.78
Venezuela	6.22	11.56	82.22
Denmark	6.22	7.11	86.67
Romania	4.89	8.00	87.11
Spain	4.89	8.22	86.89
The Philippines	4.67	8.22	87.11
Portugal	4.22	8.00	87.78
Poland	4.22	8.44	87.33
Lithuania	4.00	8.00	88.00
Ukraine	3.78	8.67	87.56

Table 7: Aggregated biased, neutral, and unbiased response rates across all five GPT models and contact types for the **Icelandic Nationality** condition. Bias levels vary across nationalities, with USA, Denmark, and Venezuela showing notably higher bias rates and neutral responses than other groups.

Religion	Biased (%)	Neutral (%)	Unbiased (%)
Russian Orthodox	22.9	10.4	66.7
Jehovah's Witness	16.0	15.1	68.9
Christian	13.1	9.8	77.1
Buddhist	6.9	8.9	84.2
No Religion	8.1	0.0	91.9
Islam	6.4	9.3	84.2

Table 8: Aggregated biased, neutral, and unbiased response rates across all five GPT models and contact types for the **Icelandic Religion** condition. Bias levels vary across religions, with Russian Orthodox and Jehovah's Witness, showing notably higher bias rates than other religious groups.

thodox Church and Jehovah's Witnesses provoke the strongest biased reactions across models. Elevated neutrality for *Jehovah's Witnesses* (15.1%) and *Russian Orthodox* (10.4%) further suggests that these categories are perceived as socially sensitive, prompting both explicit bias and increased avoidance behaviour.

In contrast, *Buddhist* and *Islam* show the lowest explicit bias rates (6.9% and 6.4%, respectively) and the highest proportions of unbiased responses (84.2%). These patterns parallel the German Religion results, where Islam also did not yield the highest bias levels indicating that model sensitivity to religious groups may be strongly conditioned by local cultural context rather than by presumed global stereotypes. Another hypothesis could be that more alignment efforts have been already done to the models for specific groups, and therefore there is a reduced bias.

Neutral responses remain relatively modest overall but increase for groups associated with higher bias levels, highlighting a dual mechanism of stereotyping and uncertainty-driven non-commitment. Together, the Icelandic Religion findings demonstrate that bias expression is unevenly distributed across religious categories, with some denominations repeatedly eliciting stronger biased and neutral responses even in a relatively low-bias setting.

5 Discussion

Cross-Condition Comparison. Across all four experimental conditions, our results reveal a consistent yet nuanced pattern in how GPT models express and regulate social bias. Explicit bias is lowest and most stable in the **German Nationality** condition, where nearly all models exhibit near-ceiling unbiased responses and minimal neutrality. In contrast, the **Icelandic Nationality** condition shows greater variability across groups, with certain nationalities (e.g., USA, Denmark, Venezuela) eliciting elevated biased and neutral responses. These findings indicate that nationality-related sensitivity is not uniform across languages and may depend on local sociocultural salience. A similar asymmetry appears across the two religion conditions. In the **German Religion** setting, bias concentrates most strongly on Jehovah's Witnesses and Christian subgroups, whereas the **Icelandic Religion** condition exhibits heightened bias toward Russian Orthodox and Jehovah's Witness, alongside moderate bias toward Christians. Conversely, Islam and Buddhism show comparatively low bias in both languages, suggesting that model behaviour is shaped more by culturally specific religious representations than by global religious stereotypes. Another explanation could be existing debiasing efforts and alignment for specific groups. Notably, neutrality increases systematically in the conditions and groups that elicit higher bias most prominently in Icelandic Religion and Icelandic Nationality consistent with a safety-driven avoidance strategy observed most strongly in GPT-5. Taken together, these results demonstrate that bias expression in current LLMs is sensitive to both linguistic context and group-specific framing, with nationality and religion showing distinct cross-linguistic patterns that underscore the importance of multilingual evaluation when assessing fairness and model alignment.

Cross-Linguistic Generalizability. The present findings demonstrate that contact-based probing extends robustly beyond English, supporting the cross-cultural validity of contact theory in computational settings (Allport, 1954; Wright et al., 1997; Amichai-Hamburger and McKenna, 2006).

Across both German and Icelandic conditions, positive contact consistently reduced explicit bias, while negative contact amplified it, mirroring prior English-only results (Raj et al., 2024). This convergence indicates that the underlying socio-cognitive mechanisms captured by contact framing might not be language-specific but might generalize across typologically related European languages. Following previous work in the field (Raj et al., 2024), the concept of cross-cultural validity is approximated by the languages in our work.

Language-Specific Patterns. Despite this generalizability, the models displayed clear nationality-specific patterns. German nationality cues elicited highly stable behaviour across model generations, with consistently low bias and minimal neutrality even under negative contact. In contrast, Icelandic nationality elicited substantially more variability, stronger sensitivity to negative framing, and elevated neutral responses especially in later-generation models such as GPT-4 and GPT-5. These cross-linguistic differences underscore the importance of culturally contextualized descriptors and social salience in LLM bias expression (Smith et al., 2022; Parrish et al., 2022), and caution against relying solely on English-centric evaluations when assessing model fairness (Bender et al., 2021).

Model Evolution. The progression from GPT-3.5 to GPT-4o reflects a clear downward trend in overt bias, consistent with claims about advances in dataset curation, alignment, and reinforcement learning from human feedback (OpenAI et al., 2023). However, GPT-5 introduced a distinct behavioural profile characterized by high rates of neutral responses, particularly in sensitive contexts (e.g., Icelandic nationality and Icelandic religion). This shift suggests that newer models increasingly rely on safety-driven non-commitment when uncertainty or social risk is detected, rather than resolving bias through more robust reasoning. Residual bias and avoidance behaviours therefore persist in nuanced forms (Zhao et al., 2023), even in the most recent model generations.

Mitigation Implications. The consistency of contact effects across languages indicates that contact-based prompt strategies (Raj et al., 2024) may serve as a practical, model-agnostic mitigation mechanism, complementing algorithmic and representational debiasing approaches (Zhang et al., 2018; Zhao et al., 2018). However, the elevated neutrality in GPT-5 suggests that mitigation efforts should also address avoidance-based failure modes, which differ qualitatively from explicit stereotyping and may obscure underlying decision-making limitations.

Limitations and Future Work. This study is limited to two languages, two dimensions of social identity, and GPT-family models. Additionally, the response space is restricted to categorical judgments, which may mask subtler forms of bias or reasoning. Also the manual review process by native speakers with human judgment might have impacted the classification of unambiguous answers. Future work should expand to more languages and cultural contexts, incorporate open-source and smaller models, and explore fine-tuning or prompting interventions that leverage contact framing more systematically (Raj et al., 2024). Longitudinal analyses across model updates may further clarify whether neutrality reflects genuine caution or unresolved representational gaps.

6 Conclusion

This work extends contact-hypothesis probing beyond earlier German-only studies by introducing a parallel Icelandic dataset for both nationality and religion, enabling a controlled cross-linguistic evaluation of bias in contemporary GPT models. Across all conditions, the models exhibit clear contact-aligned shifts positive contact reliably increases acceptance, whereas negative contact amplifies biased responses confirming the psychological predictions of contact theory (Allport, 1954; Wright et al., 1997; Raj et al., 2024). At the same time, the magnitude and distribution of bias differ systematically between languages and social categories. Nationality prompts in German yield near-ceiling unbiased responses, whereas Icelandic nationality prompts show higher variability and a greater tendency toward safety-driven neutrality. In the religion conditions, bias concentrates on some groups (e.g., Russian Orthodox, Jehovah's Witness), while Islam and Buddhism receive comparatively low bias across both languages, highlighting strong cultural and contextual effects (Smith et al., 2022; Parrish et al., 2022).

Together, these results demonstrate that bias in LLMs cannot be meaningfully assessed in a single language or cultural setting. Instead, multilingual and context-sensitive evaluation is essential

to understanding the sociolinguistic contours of model behaviour. The consistent alignment with contact framing across languages further suggests that theory-grounded approaches from social psychology can inform more reliable and generalizable bias-mitigation strategies.

Future work should extend this framework to additional languages, model families, and social categories, and explore contact-aware training or prompting as a structured path toward safer and culturally adaptive language models.

7 Acknowledgments

This work is part of the Europe Horizon project BIAS, grant agreement number 101070468, funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

Gordon W Allport. 1954. The nature of prejudice. *Addison-Wesley*.

Yair Amichai-Hamburger and Katelyn Y A McKenna. 2006. The contact hypothesis reconsidered: Interacting via the internet. *Journal of Computer-Mediated Communication*, 11(3):825–843.

Samee Arif, Zohaib Khan, Maaidah Kaleem, Suhaib Rashid, Agha Ali Raza, and Awais Athar. 2024. With a grain of salt: Are llms fair across social dimensions? *arXiv preprint arXiv:2410.12499*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Koulako Bala Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.

Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, and Tijl De Bie. 2024. Large language models reflect the ideology of their creators. *arXiv preprint arXiv:2410.18417*.

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

Lance Calvin Lim Gamboa, Yue Feng, and Mark Lee. 2025. Social bias in multilingual language models: A survey. *arXiv preprint arXiv:2508.20201*.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 122–133.

Shulin Huang, Linyi Yang, and Yue Zhang. 2025. Mceval: A dynamic framework for fair multilingual cultural evaluation of llms. *arXiv preprint arXiv:2507.09701*.

Catherine Ikae and Mascha Kurpicz-Briki. 2025. Measuring bias in german prompts to gpt models using contact hypothesis. In *AIMMES Workshop, AI Fairness Cluster Conference*.

Minsung Kim and Sanghoon Baek. 2024. Exploring large language models on cross-cultural values in connection with training methodology. *arXiv preprint arXiv:2412.08846*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian

Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Qianying Liu, Katrina Qiyao Wang, Fei Cheng, and Sadao Kurohashi. 2025. Assessing large language models in agentic multilingual national bias. *arXiv preprint arXiv:2502.17945*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-3.5 (chatgpt model). https://platform.openai.com/docs/models/gpt-3-5. Accessed 2025.

OpenAI. 2024. Gpt-4o. https://openai.com/index/hello-gpt-4o. Accessed 2025.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. 2023. Gpt-4 technical report. Technical report, OpenAI. ArXiv:2303.08774.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Eric Michael Smith, Mitchell Hall, Melanie Kambadur, Edoardo Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9180–9211.

University of Amsterdam ILLC. 2024. Multilinguality and multiculturalism: Towards more effective and inclusive neural language models. https://eprints.illc.uva.nl/id/eprint/2347/.

Stephen Wright, Arthur Aron, Tracy McLaughlin-Volpe, and Stacy Ropp. 1997. The extended contact effect: Knowledge of cross-group friendships and prejudice. *Journal of Personality and Social Psychology*, 73(1):73–90.

Pardis Sadat Zahraei and Ehsaneddin Asgari. 2025. I am aligned, but with whom? mena values benchmark for evaluating cultural alignment and multilingual bias in llms.

Brian Hu Zhang, Benjamin Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 335–340.

Jieyu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. 2023. Gptbias: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 15–20.

A Scenario Comparison

Tables 9 and 10 show the bias distribution among the different scenarios for the experiments with the German language.

Tables 11 and 12 show the bias distribution among the different scenarios for the experiments with the Icelandic language.

Across languages and identity dimensions, scenario effects followed a consistent structural pattern but differed substantially in magnitude. Workplace

Scenario	Biased (%)	Neutral (%)	Unbiased (%)
Workplace	14.9	10.2	74.9
Healthcare	4.0	4.3	91.7
Sports	3.8	9.2	87.0
Education	1.9	5.6	92.5
Community	0.2	2.7	97.1

Table 9: Bias distribution across German religion scenarios aggregated over models and contact conditions (balanced scenario counts).

Scenario	Biased (%)	Neutral (%)	Unbiased (%)
Education	3.5	2.4	94.1
Community	3.3	2.8	93.9
Sports	3.2	2.8	94.0
Healthcare	3.0	2.5	94.5
Workplace	2.6	3.1	94.4

Table 10: Bias distribution across German nationality scenarios aggregated over models and contact conditions (balanced scenario counts).

contexts were the most bias-prone across all conditions, whereas Community scenarios were consistently the most robust. Religion elicited higher bias levels than nationality in both languages, with Icelandic religion showing the strongest effects overall (26.1%), followed by German religion (14.9%). Nationality prompts produced comparatively lower bias, particularly in German (<= 3.5% across all scenarios). While the ranking of scenarios remained stable cross-linguistically, bias intensity was clearly modulated by both language and identity type.

Scenario	Biased (%)	Neutral (%)	Unbiased (%)
Workplace	26.1	13.9	60.0
Healthcare	11.5	7.8	80.7
Sports	10.7	12.0	77.2
Education	7.8	7.0	85.2
Community	4.8	7.2	88.0

Table 11: Bias distribution across Icelandic religion scenarios aggregated over models and contact conditions (balanced scenario counts).

Scenario	Biased (%)	Neutral (%)	Unbiased (%)
Workplace	11.7	13.2	75.1
Healthcare	7.4	7.0	85.6
Sports	4.2	9.3	86.4
Education	1.7	6.8	91.6
Community	0.1	6.3	93.6

Table 12: Bias distribution across Icelandic nationality scenarios aggregated over models and contact conditions (balanced scenario counts).