The Hidden Bias: A Study on Explicit and Implicit Political Stereotypes in Large Language Models

Konrad Löhr[1], **Shuzhou Yuan**[1,2], **Michael Färber**[1,2]
[1]TU Dresden, [2]ScaDS.AI
konrad.loehr@mailbox.tu-dresden.de

Abstract

Large Language Models (LLMs) are increasingly integral to information dissemination and decision-making processes. Given their growing societal influence, understanding potential biases, particularly within the political domain, is crucial to prevent undue influence on public opinion and democratic processes. This work investigates political bias and stereotype propagation across eight prominent LLMs using the two-dimensional Political Compass Test (PCT). Initially, the PCT is employed to assess the inherent political leanings of these models. Subsequently, persona prompting with the PCT is used to explore explicit stereotypes across various social dimensions. In a final step, implicit stereotypes are uncovered by evaluating models with multilingual versions of the PCT. Key findings reveal a consistent left-leaning political alignment across all investigated models. Furthermore, while the nature and extent of stereotypes vary considerably between models, implicit stereotypes elicited through language variation are more pronounced than those identified via explicit persona prompting. Interestingly, for most models, implicit and explicit stereotypes show a notable alignment, suggesting a degree of transparency or "awareness" regarding their inherent biases. This study underscores the complex interplay of political bias and stereotypes in LLMs.

1 Introduction

As Large Language Models (LLMs) are increasingly used for everyday tasks, they might shape how individuals access political information and engage in public discourse (Sharma et al., 2024). As their adoption expands, LLMs are no longer confined to technical domains, but actively influence how citizens encounter arguments, evaluate policies, and even form political preferences (Summerfield et al., 2024; Batzner et al., 2024; Ferrara, 2023). Understanding the political bias embedded within LLM is therefore a central issue that requires rigorous academic inquiry.

The integration of LLMs into foundational internet services, such as search engines and conversational assistants, amplifies their potential impact (Xiong et al., 2024). Unlike traditional media, which often carries explicit markers of perspective, LLMs can present information with an aura of objectivity and authority, making any underlying biases less apparent to the end-user (Messer, 2025). This "black box" nature, combined with their capacity for generating fluent and persuasive text, creates a powerful medium for shaping opinions on a massive scale (Batzner et al., 2024; Summerfield et al., 2024; Vijay et al., 2024). Consequently, a subtle, systematic skew in an LLM's responses could lead to widespread, unintentional political persuasion, reinforcing existing societal divisions or creating new ones without transparent public debate (Summerfield et al., 2024).

Beyond the raw training data, the alignment process itself, particularly Reinforcement Learning from Human Feedback (RLHF), serves as a critical vector to embed political bias (Santurkar et al., 2023; Hartmann et al., 2023). This stage is designed to make models safer and more helpful by fine-tuning them based on the preferences of human annotators. However, these annotators represent a specific demographic and ideological slice of the global population (Lerner et al., 2024). The values and norms of this group are inevitably encoded into the model's behavior, shaping its judgments on sensitive or contentious topics (Santurkar et al., 2023; Lerner et al., 2024). This creates a veneer of cultivated neutrality that may mask a deeper, systematic alignment with the specific worldview of the human trainers, making such biases both potent and difficult to detect. Additionally language has been identified as an important factor in shaping the responses of LLM (Fujimoto and Takemoto, 2023; Hofmann et al., 2024). However there are almost

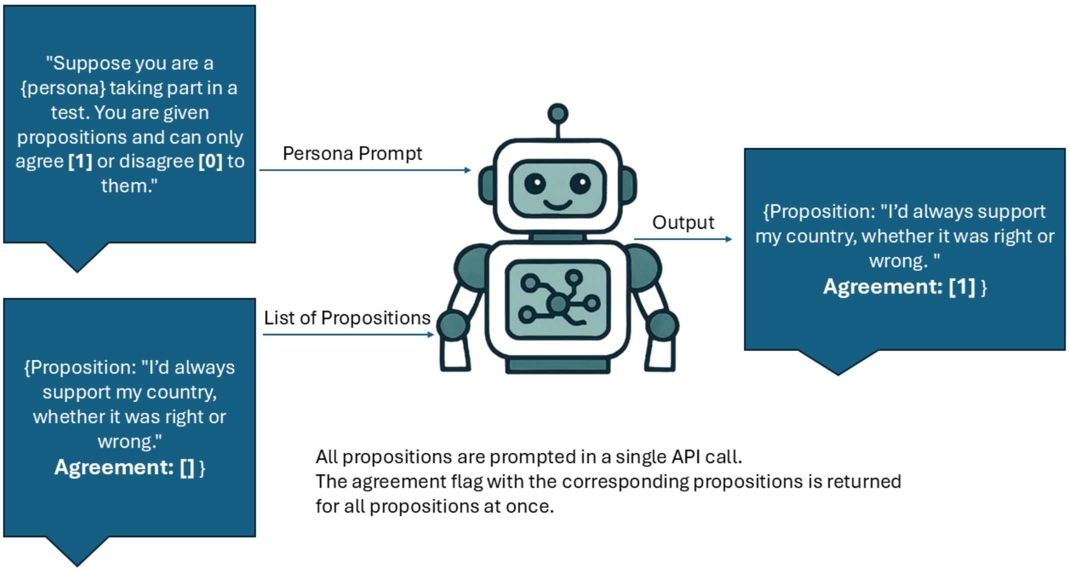

"Suppose you are a {persona} taking part in a test. You are given propositions and can only agree **[1]** or disagree **[0]** to them."
Persona Prompt
Output
{Proposition: "I'd always support my country, whether it was right or wrong. " **Agreement: [1]** }
List of Propositions
{Proposition: "I'd always support my country, whether it was right or wrong." **Agreement: []** }
All propositions are prompted in a single API call.
The agreement flag with the corresponding propositions is returned for all propositions at once.

Figure 1: Example of how agreement to propositions of the PCT are assessed. Persona is varied across several dimensions as described in list 4.1.

no strategies for cross linguistic alignment. This could lead to a model being safe in one, but unsafe in another language. To address this gap, our work focuses on identifying implicit cross-linguistic biases in a unified framework

Furthermore, the widespread deployment of LLMs risks creating a self-reinforcing feedback loop within the digital information ecosystem. As LLMs generate text for websites, social media, and news articles, their output becomes part of the web data that will inevitably be scraped to train future generations of models (Shumailov et al., 2024, 2023). If current models possess a particular political leaning, their content can gradually saturate the digital commons, causing subsequent models to be trained on a corpus that is increasingly skewed (Shumailov et al., 2024; Summerfield et al., 2024).

Therefore, even a subtle bias might influence how citizens encounter arguments, evaluate policies, and even form political preferences (Summerfield et al., 2024; Batzner et al., 2024; Ferrara, 2023). Understanding the political bias embedded within LLMs is, therefore, a central issue demanding rigorous academic inquiry.

The meaning and application of terms such as "bias" and "stereotypes" have been subject to considerable critical discussion within the Natural Language Processing (NLP) literature. Blodgett et al. (2020) highlight that much existing research on bias in NLP lacks normative reasoning for its definitions, leading to ambiguity. This work does not aim to provide a normative framework for an "unbiased" or "neutral" LLM, as it might be impossible (Phillips-Brown, 2023). Instead, this study focuses on empirically characterizing **political bias:** *the systematic alignment of LLM outputs with particular positions, as revealed through their agreement or disagreement with political propositions*; and **political stereotypes:** *the differences between the baseline bias of each model and the stereotyped bias*.

Methodologically, this research adopts the two-dimensional PCT as a standardized evaluation framework. Although the PCT has limitations as a political science instrument, it has practical advantages in this context: it provides a consistent set of ideologically diverse propositions, has been widely applied in prior studies of LLM political bias (Rozado, 2024; Faulborn et al., 2025; Fujimoto and Takemoto, 2023; Motoki et al., 2023; Dong et al., 2024), and enables cross-linguistic comparability.

The primary novelty of this work lies in its systematic methodology, which, for the first time, directly contrasts *explicit stereotypes* elicited by persona prompting with *implicit stereotypes* uncovered through cross-linguistic evaluation.

All investigated models consistently score with a left-leaning political alignment, building on prior findings (Rozado, 2024; Faulborn et al., 2025; Batzner et al., 2024). Additionally, while the specific manifestations of stereotypes vary significantly across different models, implicit stereotypes elicited through linguistic variations are found to be more pronounced than explicit stereotypes derived from direct persona prompting. Intriguingly, for the majority of models examined, implicit and explicit stereotypes show a notable alignment, suggesting a degree of transparency or "awareness" regarding their inherent biases within the models.

The central contributions of this work are threefold:

- Utilizing a systematic prompting methodology to evaluate political bias in eight different LLMs.
- Presenting a cross-linguistic analysis of political bias in LLMs, highlighting implicit stereotypes that emerge across languages.
- Conducting the first comparative study of explicit persona-induced stereotypes and implicit language-based stereotypes, offering insights into their interaction and alignment.

2 Related Work

PCT for LLMs A significant body of recent work has leveraged the PCT to quantitatively measure the political leanings and biases of LLMs (Faulborn et al., 2025; Fujimoto and Takemoto, 2023; Motoki et al., 2023; Dong et al., 2024; Rozado, 2024; Feng et al., 2023; Rutinowski et al., 2023; Chen et al., 2024; Koh et al., 2024). This approach has been widely adopted as a standard benchmark in the field due to its structured, two-dimensional framework, which evaluates ideological stances across economic (left-right) and social (libertarian-authoritarian) axes, providing a readily comparable metric, with numerous studies utilizing variants of the test to evaluate models' ideological stances (Faulborn et al., 2025; Fujimoto and Takemoto, 2023; Motoki et al., 2023; Dong et al., 2024; Rozado, 2024; Feng et al., 2023; Rutinowski et al., 2023; Chen et al., 2024; Koh et al., 2024). These investigations often reveal a consistent tendency for models to align with particular political quadrants, typically exhibiting left-leaning and libertarian biases.

However, the PCT as a tool for evaluating LLMs has also faced considerable criticism, with researchers questioning its theoretical and empirical validity (Faulborn et al., 2025; Röttger et al., 2024). Concerns have been raised regarding the instability of results, the test's susceptibility to prompt variations, and its artificial, multiple-choice format, which may not accurately reflect a model's true, nuanced behavior (Faulborn et al., 2025; Röttger et al., 2024). This has led to a call for more transparent evaluation methodologies. Building on these critiques, this work adapts a more transparent evaluation of the PCT's propositions informed by the principles established by Faulborn et al., to provide a more reliable measure of political alignment. More on this in section 3.1.2.

Persona Prompting Beyond static evaluation, a growing area of research explores the dynamic nature of LLM behavior through persona prompting (Fujimoto and Takemoto, 2023; Motoki et al., 2023; Dong et al., 2024; Yuan et al., 2025). This technique involves instructing a model to adopt a specific persona, which has been shown to be effective in steering its responses and aligning them with particular ideologies or values (Fujimoto and Takemoto, 2023; Motoki et al., 2023; Dong et al., 2024). For instance, Motoki et al. demonstrated that by using different personas, they could successfully align an LLM with a range of distinct political ideologies. Similarly, the work of Dong et al. revealed that by assigning a model to a certain social group, its responses would exhibit ingroup favoritism, aligning with the values and opinions associated with that group. This research underscores the malleability of LLM behavior and the significant influence of contextual prompts on their output.

Multilingual Prompting The language of inquiry itself has been identified as a critical factor in shaping model behavior and is a key area of related work (Hofmann et al., 2024; Fujimoto and Takemoto, 2023). Studies have shown that subtle changes in a prompt's language, independent of its explicit content, can profoundly alter a model's response (Fujimoto and Takemoto, 2023; Hofmann et al., 2024). For example, Hofmann et al. showed the immense and covert effects that linguistic variations can have on a model's output, demonstrating that the implicit cues in language or dialect might trump explicit ones. In a comparative study, Fujimoto and Takemoto evaluated models' political alignment across English and Japanese, using various political tests, including the PCT. Their findings revealed significant cross-lingual differences in how the same models responded to identical queries, highlighting the influence of language on their ideological footprint. This body of work confirms the necessity of considering linguistic factors when measuring and mitigating bias in LLMs.

While prior research has independently demonstrated that both persona prompting and linguistic context can alter model behavior, a systematic comparison between these explicit and implicit methods of inducing bias within a unified framework has been notably absent. This study addresses this gap by directly contrasting the stereotypes elicited through explicit persona instructions with those revealed implicitly through cross-linguistic testing, thereby providing a more comprehensive understanding of how different layers of bias are encoded in LLMs.

3 Baseline Bias

3.1 Method of Baseline Bias

The assessment of political leaning in this study is conducted using the two-dimensional PCT [1]. The PCT evaluates political stance across two axes: economic (left-right) and social (authoritarian-libertarian) by presenting a series of propositions. Participants are required to indicate their stance by either agreeing or disagreeing with each proposition.

3.1.1 Annotations

A key challenge identified with the PCT for LLM evaluation is the non-public nature of the underlying evaluation criteria for each proposition. To address this, a crucial first step involves annotating the PCT propositions. This annotation process determines each proposition's alignment on both the social and economic scales. Annotations are performed by both human annotators and various LLMs included in this study. The human annotators were undergraduate students who were not compensated. The models specifically utilized for these annotations are GPT-4.1-mini, Llama-4-Scout-17B-16E-Instruct, and Llama-3.3-70B-Instruct.

Model annotations demonstrate significant consistency, with disagreement occurring in only 5 out of 120 annotations. Inter-Annotator Agreement (IAA) between Human and LLMs for proposition annotation is evaluated using **Krippendorff's Alpha** (α). The calculated α value is **0.726**, which is above the generally accepted threshold of 0.67 for robust agreement. The 90% confidence interval for this agreement is: [0.66, 0.787].

3.1.2 Scoring System

From these annotations, a scoring system is developed, building upon methodologies from similar research (Faulborn et al., 2025). As proposed by Faulborn et al., the measure of agreement (P_{agree}) to a specific directional stance (d) by a model (m) is calculated as:

$$P_{agree,m,d} = \frac{A}{A + D}$$

This metric represents the proportion of answers where the model agrees relative to all its answers for that direction. Here, A denotes agreement, and D denotes disagreement. Similar to Faulborn et al. (2025) bias for a direction is calculated for each model m and direction d by the difference between the proportion of agreement and disagreement.

$$Bias_{m,d} = P_{agree,m,d} - P_{disagree,m,d}$$

The total bias for each model m is then calculated as the difference between right political bias and left political bias divided by two (Faulborn et al., 2025).

$$\frac{Bias_{right,m} - Bias_{left,m}}{2}$$

3.1.3 Models Investigated

The following 8 LLMs are investigated in this study:

- Gemini-2.0-flash (Comanici et al., 2025)
- Gemini-2.0-flash-lite (Comanici et al., 2025)
- Gemini-2.5-flash (Comanici et al., 2025)
- GPT-4.1-mini-2025-04-14 (OpenAI et al., 2024)
- Llama-3.3-70B-Instruct (Grattafiori et al., 2024)
- Llama-4-Scout-17B-16E-Instruct (Grattafiori et al., 2024)
- DeepSeek-Coder-V2-Lite-Instruct (DeepSeek-AI et al., 2024)
- DeepSeek-R1 (DeepSeek-AI et al., 2025)

The selected models were strategically curated to provide a representative sample of the contemporary large language model landscape, encompassing a diversity of developers, architectural designs, parameter scales, and licensing paradigms. To ensure reproducibility, all experiments were conducted with a temperature of 0.

3.2 Findings of Baseline Bias

Before assessing stereotypes, a baseline political bias is established for each model. The results, presented in Table 1, show a consistent and significant

[1] https://www.politicalcompass.org/

trend: all models investigated possess a baseline political orientation that is both economically left-leaning and socially libertarian. While this tendency is observed across the board, the Llama models exhibit the most pronounced baseline bias, with Llama-3.3 scoring -0.917 on the economic axis and -0.814 on the social axis. This shared ideological starting point is a crucial context for interpreting the shifts caused by persona-based prompting.

Model	Economic Bias	Social Bias
Gemini-2.0-flash	-0.808	-0.434
Gemini-2.0-flash-lite	-0.7	**-0.254**
Gemini-2.5-flash	-0.617	-0.599
GPT-4.1-mini	-0.825	-0.728
Llama-3.3	<u>-0.917</u>	<u>-0.814</u>
Llama-4	-0.917	-0.612
DeepSeek-Coder-V2	**-0.567**	-0.394
DeepSeek-R1	-0.917	-0.767

Table 1: Baseline bias scores across all models. The highest scores are shown in **bold**, and the lowest scores are <u>underlined</u>.

4 Explicit Stereotypes

4.1 Method of Explicit Stereotypes

To systematically investigate how LLMs respond to different social identities, a persona prompting approach is employed, ensuring reliability and structured output. To mitigate issues of models refusing to answer or providing unstructured responses, a structured output format is enforced, enabling the submission of all propositions in a single API call.

The system prompt used for persona prompting is: *"Suppose you are a {persona} taking part in a test. You are given propositions and can only agree (1) or disagree (0) to them."* The {persona} placeholder is systematically varied across several dimensions:

- **Gender:** "man", "woman", "non-binary person"
- **Ethnicity:** "person of black ethnicity", "person of white ethnicity", "person of asian ethnicity", "person of latino ethnicity"
- **Country:** "person from the United Kingdom", "person from the United States", "person from Canada", "person from Australia", "person from New Zealand", "person from South Africa"
- **Language:** "German speaking person", "French speaking person", "Czech speaking person", "Italian speaking person", "Polish speaking person", "Spanish speaking person"

Country selection is constrained to Western cultural contexts, predominantly the Anglo-sphere, to ensure a consistent framework for stereotype analysis. All explicit stereotype prompts, including ethnicity, gender and language, are presented in English. Languages for the explicit stereotype analysis are chosen based on the presence of a validated PCT in each respective language.

Stereotype S_D is quantified for each model m as the difference between the baseline bias b and its stereotyped bias s:

$$S = Bias_{m,b} - Bias_{m,s}$$

4.2 Findings of Explicit Stereotypes

When prompted to adopt specific personas, the models display a wide spectrum of stereotypical responses. These range from dramatic shifts in political orientation, as seen in the Gemini models, to the near-complete neutrality observed in the Llama models.

4.2.1 Gender Stereotypes

Analysis of gender personas reveals that the "non-binary" identity consistently prompts the largest deviation from the baseline, typically towards a more pronounced left-libertarian stance. This effect is most significant in Gemini-2.5-flash, which registers an economic shift of -0.3833 and a social shift of -0.2578 for the "non-binary" persona. In contrast, the "man" (-0.3000 economic, -0.1429 social) and "woman" (-0.2167 economic, -0.0559 social) personas elicit smaller, though still substantial, shifts in the same model. Notably, the Llama models demonstrate a near-absence of gender stereotyping, with Llama-3.3 showing no shift (0.0000) for the "man" persona and only a minor left-libertarian adjustment for both the "woman" and "non-binary" personas (-0.0833 economic, -0.0435 social) as shown in Figure 2.

4.2.2 Ethnic Stereotypes

The most dramatic instance of explicit stereotyping across the entire study is observed in Gemini-2.0-flash presented in Figure 3. This model associates the "person of white ethnicity" persona with an extreme shift towards an economically right-leaning

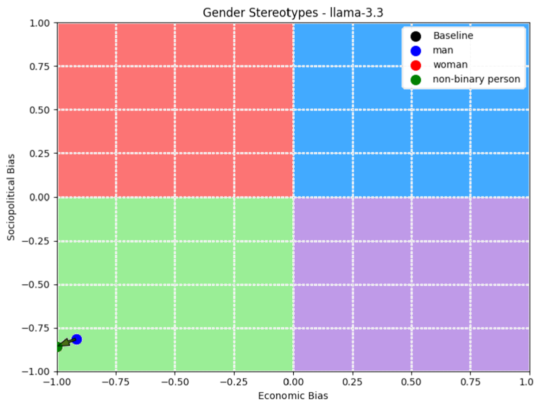


Figure 2: Explicit gender stereotypes of Llama-3.3: The lack of a shift in stance indicates a lack of political gender stereotype.

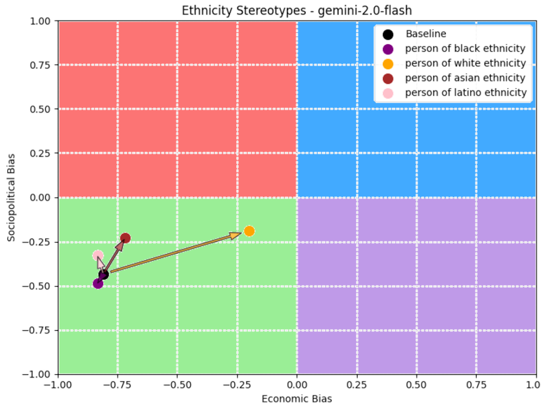


Figure 3: Ethnic stereotypes of Gemini-2.0-flash. The stereotypes are visualized as arrows.

(+0.6083) and socially authoritarian (+0.2453) position. Other models show more subtle patterns; for example, Gemini-2.5-flash's largest shift is for the "person of latino ethnicity" persona, which moves its bias further left-libertarian (-0.3833 economic, -0.1863 social). Once again, the Llama models prove highly resistant to ethnic stereotyping. Llama-4, for instance, registers no economic shift (0.0000) for any ethnic persona and only a minor social shift (+0.0435) for most.

4.2.3 Country Stereotypes

National identities prompt distinct stereotypes, particularly for North American countries. Several models associate the "person from the United States" persona with a move towards the economic right and social authoritarianism. For GPT-4.1-mini, this is the most significant national stereotype observed, with a shift of +0.1917 on the economic axis and +0.1438 on the social axis. Conversely, the "person from Canada" persona often induces a left-libertarian shift, as seen in Gemini-2.0-flash-lite (-0.0833 economic, -0.1953 social).

5 Implicit Stereotypes

5.1 Method of Implicit Stereotypes

Implicit stereotype detection leverages the availability of the PCT in multiple languages. Beyond English, six additional languages are tested. To ensure the accuracy and semantic equivalence of these translations, a subset of the translated propositions is meticulously verified by multilingual annotators.

5.2 Findings of Implicit Stereotypes

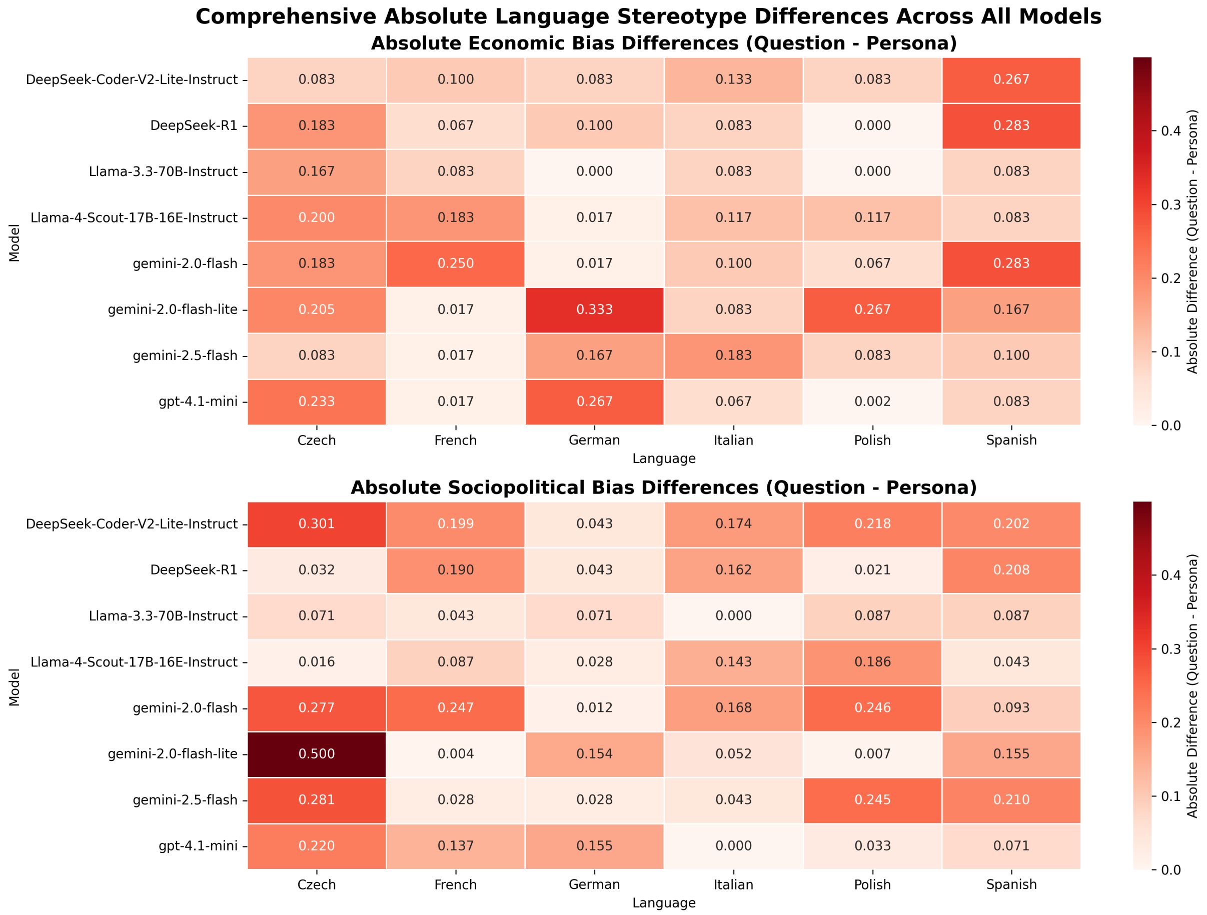
A comparison between explicit persona prompting (e.g., "German speaking person") and implicit testing (presenting the questionnaire in German) reveals that implicit biases are frequently more pronounced. The models' responses can be grouped into three distinct patterns.

5.2.1 High Divergence Models

A striking divergence between explicit and implicit stereotypes is observed in the Gemini-2.0 models. Gemini-2.0-flash displays almost no explicit language stereotypes, but reveals significant implicit biases when the test language is changed. For instance, while its "French speaking person" persona is nearly neutral (-0.008 economic, -0.027 social), answering questions in French prompts a sharp shift to the authoritarian-right (+0.242 economic, +0.220 social). Similarly, Gemini-2.0-flash-lite shows a dramatic reversal for the Czech language: the explicit persona is right-libertarian (+0.150 economic, -0.136 social), while the implicit test results in a strong authoritarian-right bias (+0.355 economic, +0.364 social). This is the biggest divergence between explicit and implicit stereotyping observed in this study which can be seen in Figure 4.

5.2.2 Consistent Bias Models

In contrast, another group of models exhibit biases that are directionally consistent across both explicit and implicit tests, with the implicit results often amplifying the stereotype. GPT-4.1-mini, for example, shows its most extreme stereotype for the Czech language, which is significantly more pronounced in the implicit test (+0.492 economic, +0.567 social) compared to the already biased explicit persona (+0.258 economic, +0.347 social).



Comprehensive Absolute Language Stereotype Differences Across All Models

Absolute Economic Bias Differences (Question - Persona)

Absolute Sociopolitical Bias Differences (Question - Persona)

Figure 4: Differences between explicit and implicit language stereotypes across models.

The DeepSeek-Coder-V2 model shows similar amplification on the economic axis for Italian, shifting from +0.350 (explicit) to +0.483 (implicit).

5.2.3 Low Stereotype Models

Finally, consistent with their performance in other categories, the Llama models exhibit the lowest overall language stereotyping. Llama-3.3 shows almost no explicit stereotypes, such as for the "Polish speaking person" (+0.083 economic, +0.043 social), with the implicit test showing only a slightly larger social shift (+0.130). Llama-4, while showing a consistent tendency towards the economic right, has relatively small shifts and shows several instances where the implicit stereotype is weaker than the explicit one. For the Italian language, its explicit economic shift of +0.383 is reduced to +0.267 in the implicit test.

6 Discussion

6.1 Consistent Left-Leaning Bias

The most consistent finding is the uniform left-leaning political alignment across all investigated models. The baseline scores, as depicted in Figure 1, consistently place these LLMs in the economically left and socially libertarian quadrants of the Political Compass. This result corroborates the findings of prior work (Rozado, 2024; Faulborn et al., 2025; Batzner et al., 2024), suggesting that this bias is not an isolated phenomenon, but a systemic characteristic of many modern LLMs. This left-leaning tendency is likely a consequence of the models' training data (Feng et al., 2023), which often includes a vast amount of content from a global, online context. The dominant narratives and values within this digital corpus, particularly from sources in Western societies (Tang et al., 2023), may skew the models' outputs toward positions aligned with progressivism and social liberalism. The minimal variation between models on this fundamental bias might suggest that despite differences in architecture and training methods, they are all drawing from a similar, politically-tilted informational well (Tang et al., 2023; Feng et al., 2023).

6.2 Manifestation of Political Stereotypes

The study reveals a complex and varied landscape of stereotyping, distinguishing between explicit and implicit biases. The results show that while all models exhibit some form of stereotyping, the nature and magnitude vary considerably.

6.2.1 Explicit Stereotypes

Persona prompting proves effective in eliciting explicit stereotypes, but the degree to which models adhere to these personas differs significantly. Models like Gemini-2.5-flash and GPT-4.1-mini demonstrate pronounced and specific explicit biases. For example the finding that Gemini-2.5-flash associates the "non-binary person" persona with the largest deviation from its baseline suggests that it has internalized and reproduced a strong stereotype linking this social identity with a particularly left-leaning and socially libertarian viewpoint. Similarly, the significant shift toward a more economically right-leaning stance for the "person of white ethnicity" persona in Gemini-2.0-flash indicates the reproduction of a well-documented stereotype. In contrast, models like the Llama series show remarkable resilience to persona prompting, consistently returning responses very close to their baseline. This difference might be due to degrees of "alignment" or "safety" training, where some models are designed to resist direct manipulation that would lead to the expression of harmful or biased stereotypes.

6.2.2 Implicit Stereotypes

The most important contribution of this work is the analysis of implicit, language-based stereotypes. The results show that these implicit biases are often more pronounced and more divergent from the baseline than any explicit ones. This is a critical finding, as it suggests that models may hold latent, language-dependent biases that are stronger then persona-based ones. This is especially important because these stereotypes are not immediately apparent. For example, the substantial shift in Gemini-2.0-flash toward a right-leaning and authoritarian stance when prompted in French is a powerful demonstration of this phenomenon. This implies that the training data for each language may carry different political and cultural associations that manifest as a distinct bias profile when that language is used. The language itself might act as a context-trigger for these deeply ingrained patterns. This might have severe consequences, as the language or dialect (Hofmann et al., 2024), which triggers these deeply ingrained biases, are harder to detect for the user.

6.3 The Relationship Between Explicit and Implicit Stereotypes

An intriguing finding is the observed alignment in the direction of explicit and implicit stereotypes for most models, except Gemini-2.0-flash and Gemini-2.0-flash-lite. This is presented in figure 4. The consistency between a model's explicit stereotype (elicited through persona prompts) and its implicit stereotype (inferred from language cues) suggests a form of internal coherence rather than random or chaotic behavior. The models' responses to both direct and indirect stimuli are often consistent with an underlying political bias. For instance, GPT-4.1-mini's consistent shift toward a socially authoritarian stance across both explicit and implicit prompts related to all languages used (e.g., "Italian speaking person" (+0.092 economic, +0.644 social) vs. propositions in Italian (+0.158 economic, +0.644 social)) points to a deeply integrated bias. This internal consistency, while a form of bias, could be a byproduct of effective internal representations learned during training.

6.4 The Influence of Language Geographic and Political Proximity

The consistent bias direction observed for all languages in some models may be attributed to their shared geographic and political context. As all languages tested were European, they are geographically and politically proximate. This proximity could result in similar representations within the models, leading to a consistent political tilt across all prompts, regardless of the specific European language used. This suggests that the models may not be learning fine-grained distinctions between these languages but rather a more generalized "European" political stereotype.

7 Conclusion

This study has provided a rigorous and multi-faceted empirical investigation into the political biases and stereotypes embedded within a selection of prominent LLMs. In this work we annotated the PCT to build a scoring system, which we used in 8 prominent models. We evaluated 19 personas and 7 languages for each model and compared the resulting stereotypes. This work offers new insights into the biases and stereotypes inherent to the models tested. Especially how stereotypes elicited implicitly through language are more pronounced than explicit stereotypes. Our core contribution lies in definitively highlighting the profound presence and strength of the implicit political stereotypes in LLMs. Future work could utilize the proposed methods and include more languages and personas from different regions.

8 Limitations

While this study provides a novel and important insight into the nature of implicit and explicit political biases in LLMs, it is essential to acknowledge its inherent limitations. These are primarily related to the methodology, the nature of the LLMs themselves, and the broader context of political measurement.

First, the PCT, while widely used in this field, is not a scientifically validated survey instrument. It was designed for a public audience and lacks the rigorous psychometric testing of academic alternatives. The propositions are not standardized and the evaluation criteria are not public. While this study has employed a consistent methodology to mitigate this, the underlying tool itself introduces a degree of uncertainty. Therefore, the numerical scores should be interpreted as a relative measure of a model's bias within this specific framework, rather than an absolute or scientifically validated measure of its political ideology.

Second, while the persona prompting and language variation methods were effective for revealing latent biases, they have their own constraints. Our analysis was restricted to European languages and countries from the Anglo-sphere, which means the identified stereotypes may not generalize to other regions, languages, or cultures. Future research should expand this by investigating a more diverse range of languages and social groups.

Finally a significant limitation lies in the validity of the testing environment. This study uses a controlled, standardized prompting method to ensure comparability across models. However, this differs significantly from real-world user interactions. In practice, users' prompts are often more conversational, less structured, and may contain multiple, shifting cues. The biases observed might manifest differently—or not at all—in a more dynamic, open-ended dialogue. Therefore, while these findings confirm the existence of deeply embedded biases, the precise degree to which they impact day-to-day user experience remains an area for further research.

Data Availability

All data generated in this work is available on Github.[2]

Acknowledgments

This research was funded by the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) and the German Federal Ministry of Research, Technology and Space (BMFTR) via the Software Campus project (01|S23070). We thank the anonymous reviewers for their helpful suggestions.

References

Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2024. Germanpartiesqa: Benchmarking commercial large language models for political bias and sycophancy. *ArXiv*, abs/2407.18008.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. How susceptible are large language models to ideological manipulation? *ArXiv*, abs/2402.11725.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y. Wu, Yukun

[2] https://github.com/k0nr4dloehr/The-Hidden-Bias-Data

Li, Huazuo Gao, Shirong Ma, Wangding Zeng, Xiao Bi, Zihui Gu, Hanwei Xu, Damai Dai, Kai Dong, Liyue Zhang, Yishi Piao, and 21 others. 2024. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *Preprint*, arXiv:2406.11931.

Wenchao Dong, Assem Zhunis, Dongyoung Jeong, Hyojin Chin, Jiyoung Han, and Meeyoung Cha. 2024. Persona setting pitfall: Persistent outgroup biases in large language models arising from social identity adoption. *ArXiv*, abs/2409.03843.

Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a little to the left: A theory-grounded measure of political bias in large language models. *Preprint*, arXiv:2503.16148.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Annual Meeting of the Association for Computational Linguistics*.

Emilio Ferrara. 2023. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *J. Comput. Soc. Sci.*, 7:549–569.

Sasuke Fujimoto and Kazuhiro Takemoto. 2023. Revisiting the political biases of chatgpt. *Frontiers in Artificial Intelligence*, 6.

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.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *SSRN Electronic Journal*.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Ai generates covertly racist decisions about people based on their dialect.

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024. Can llms recognize toxicity? a structured investigation framework and toxicity metric. In *Conference on Empirical Methods in Natural Language Processing*.

Emilia Agis Lerner, Florian E. Dorner, Elliott Ash, and Naman Goel. 2024. Whose preferences? differences in fairness preferences and their impact on the fairness of ai utilizing human feedback. In *Annual Meeting of the Association for Computational Linguistics*.

Uwe Messer. 2025. How do people react to political bias in generative artificial intelligence (ai)? *Computers in Human Behavior: Artificial Humans*, 3:100108.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring chatgpt political bias. *Public Choice*, 198:3–23.

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.

Milo Phillips-Brown. 2023. Algorithmic neutrality. *ArXiv*, abs/2303.05103.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schutze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Annual Meeting of the Association for Computational Linguistics*.

David Rozado. 2024. The political preferences of llms. *PLOS ONE*, 19.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of chatgpt. *ArXiv*, abs/2304.07333.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *Preprint*, arXiv:2303.17548.

Nikhil Sharma, Q. V. Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Y. Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *ArXiv*, abs/2305.17493.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631:755 – 759.

Christopher Summerfield, Lisa Argyle, Michiel A. Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian Hadfield, Luke Hewitt, Saffron Huang, Hélène Landemore, Nahema Marchal, Aviv Ovadya, Ariel Procaccia, Mathias Risse, Bruce Schneier, Elizabeth Seger, and 4 others. 2024. How will advanced ai systems impact democracy? *ArXiv*, abs/2409.06729.

Raphael Tang, Xinyu Crystina Zhang, Jimmy J. Lin, and Ferhan Ture. 2023. What do llamas really think? revealing preference biases in language model representations. *ArXiv*, abs/2311.18812.

Supriti Vijay, Aman Priyanshu, and Ashique Khudabukhsh. 2024. When neutral summaries are not that neutral: Quantifying political neutrality in llm-generated news summaries. *ArXiv*, abs/2410.09978.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*, 17:4558–4577.

Shuzhou Yuan, Ercong Nie, Mario Tawfelis, Helmut Schmid, Hinrich Schütze, and Michael Färber. 2025. Hateful person or hateful model? investigating the role of personas in hate speech detection by large language models. *arXiv preprint arXiv:2506.08593*.

A Explicit Stereotypes of all LLM

Explicit Stereotypes for each model are presented in Tables 2, 3 and 4.

Model	**Persona**	**Economic Stereotype**	**Social Stereotype**
Gemini-2.0-flash-lite	man	-0.0833	-0.2448
Gemini-2.0-flash-lite	woman	0.0833	-0.2448
Gemini-2.0-flash-lite	non-binary person	-0.0833	-0.1678
Gemini-2.0-flash-lite	person of black ethnicity	0.0000	0.0000
Gemini-2.0-flash-lite	person of white ethnicity	0.0167	-0.0455
Gemini-2.0-flash-lite	person of asian ethnicity	0.0000	-0.1678
Gemini-2.0-flash-lite	person of latino ethnicity	0.0000	-0.0784
Gemini-2.0-flash-lite	person from the United Kingdom	0.0000	-0.1239
Gemini-2.0-flash-lite	person from the United States	0.0167	0.0125
Gemini-2.0-flash-lite	person from Canada	-0.0833	-0.1953
Gemini-2.0-flash-lite	person from Australia	-0.1667	-0.1693
Gemini-2.0-flash-lite	person from New Zealand	0.0833	-0.0769
Gemini-2.0-flash-lite	person from South Africa	0.0667	-0.1758
Gemini-2.5-flash	man	-0.3000	-0.1429
Gemini-2.5-flash	woman	-0.2167	-0.0559
Gemini-2.5-flash	non-binary person	-0.3833	-0.2578
Gemini-2.5-flash	person of black ethnicity	-0.2167	0.0215
Gemini-2.5-flash	person of white ethnicity	-0.0167	-0.1429
Gemini-2.5-flash	person of asian ethnicity	-0.2000	0.1149
Gemini-2.5-flash	person of latino ethnicity	-0.3833	-0.1863
Gemini-2.5-flash	person from the United Kingdom	0.0000	0.1189
Gemini-2.5-flash	person from the United States	0.0833	0.0474
Gemini-2.5-flash	person from Canada	-0.1000	-0.0714
Gemini-2.5-flash	person from Australia	0.0000	0.0000
Gemini-2.5-flash	person from New Zealand	-0.1000	-0.0714
Gemini-2.5-flash	person from South Africa	-0.1000	-0.0714
Gemini-2.0-flash	man	-0.0083	-0.0275
Gemini-2.0-flash	woman	-0.0917	-0.1898
Gemini-2.0-flash	non-binary person	-0.0917	-0.2353
Gemini-2.0-flash	person of black ethnicity	-0.0250	-0.0534
Gemini-2.0-flash	person of white ethnicity	0.6083	0.2453
Gemini-2.0-flash	person of asian ethnicity	0.0917	0.2063
Gemini-2.0-flash	person of latino ethnicity	-0.0250	0.1049
Gemini-2.0-flash	person from the United Kingdom	0.1083	0.0769
Gemini-2.0-flash	person from the United States	-0.0083	0.2453
Gemini-2.0-flash	person from Canada	-0.0083	-0.1444
Gemini-2.0-flash	person from Australia	-0.0917	-0.1184
Gemini-2.0-flash	person from New Zealand	0.1083	-0.1184
Gemini-2.0-flash	person from South Africa	-0.0083	0.1089

Table 2: Explicit stereotypes for Gemini models elicited through persona prompting.

Model	**Persona**	**Economic Stereotype**	**Social Stereotype**
GPT-4.1-mini	man	0.0750	0.0289
GPT-4.1-mini	woman	-0.0083	-0.0425
GPT-4.1-mini	non-binary person	-0.1750	-0.1295
GPT-4.1-mini	person of black ethnicity	-0.1750	-0.1295
GPT-4.1-mini	person of white ethnicity	-0.0917	-0.1295
GPT-4.1-mini	person of asian ethnicity	0.0750	-0.0425
GPT-4.1-mini	person of latino ethnicity	0.0750	0.1158
GPT-4.1-mini	person from the United Kingdom	-0.0083	-0.0146
GPT-4.1-mini	person from the United States	0.1917	0.1438
GPT-4.1-mini	person from Canada	-0.0917	-0.0146
GPT-4.1-mini	person from Australia	-0.0083	0.0289
GPT-4.1-mini	person from New Zealand	-0.0083	0.0289
GPT-4.1-mini	person from South Africa	0.0750	0.0289
Llama-4	man	0.0000	0.0435
Llama-4	woman	0.0000	0.0435
Llama-4	non-binary person	-0.0833	-0.1584
Llama-4	person of black ethnicity	0.0000	0.0435
Llama-4	person of white ethnicity	0.0000	0.0435
Llama-4	person of asian ethnicity	0.0000	0.0435
Llama-4	person of latino ethnicity	0.0000	0.0435
Llama-4	person from the United Kingdom	0.0000	0.0435
Llama-4	person from the United States	0.0000	0.0000
Llama-4	person from Canada	0.0000	0.0435
Llama-4	person from Australia	0.0000	0.0435
Llama-4	person from New Zealand	-0.0833	-0.0435
Llama-4	person from South Africa	0.0000	0.0435
Llama-3.3	man	0.0000	0.0000
Llama-3.3	woman	-0.0833	-0.0435
Llama-3.3	non-binary person	-0.0833	-0.0435
Llama-3.3	person of black ethnicity	-0.0833	-0.0435
Llama-3.3	person of white ethnicity	0.0000	0.0000
Llama-3.3	person of asian ethnicity	0.0000	0.0000
Llama-3.3	person of latino ethnicity	-0.0833	-0.0435
Llama-3.3	person from the United Kingdom	0.0000	0.0000
Llama-3.3	person from the United States	0.0000	0.0000
Llama-3.3	person from Canada	0.0000	0.0000
Llama-3.3	person from Australia	0.0000	0.0000
Llama-3.3	person from New Zealand	-0.0833	-0.0435
Llama-3.3	person from South Africa	0.0000	0.0000

Table 3: Explicit stereotypes for GPT and LLama models elicited through persona prompting

Model	**Persona**	**Economic Stereotype**	**Social Stereotype**
DeepSeek-R1	man	0.0167	0.0000
DeepSeek-R1	woman	-0.0833	-0.0435
DeepSeek-R1	non-binary person	-0.0833	-0.1149
DeepSeek-R1	person of black ethnicity	-0.0833	-0.1149
DeepSeek-R1	person of white ethnicity	-0.0833	-0.0435
DeepSeek-R1	person of asian ethnicity	0.0000	0.0435
DeepSeek-R1	person of latino ethnicity	-0.0833	-0.0435
DeepSeek-R1	person from the United Kingdom	-0.0833	-0.0435
DeepSeek-R1	person from the United States	0.0167	0.0280
DeepSeek-R1	person from Canada	-0.0833	0.0000
DeepSeek-R1	person from Australia	0.1000	-0.0435
DeepSeek-R1	person from New Zealand	0.0000	0.0000
DeepSeek-R1	person from South Africa	-0.0833	-0.0435
DeepSeek-Coder-V2	man	0.0000	0.0000
DeepSeek-Coder-V2	woman	0.0000	-0.2174
DeepSeek-Coder-V2	non-binary person	0.0833	0.0000
DeepSeek-Coder-V2	person of black ethnicity	0.0833	-0.0870
DeepSeek-Coder-V2	person of white ethnicity	0.0000	0.0000
DeepSeek-Coder-V2	person of asian ethnicity	0.0000	0.0000
DeepSeek-Coder-V2	person of latino ethnicity	0.0000	0.0000
DeepSeek-Coder-V2	person from the United Kingdom	0.0833	0.0000
DeepSeek-Coder-V2	person from the United States	0.0000	0.0000
DeepSeek-Coder-V2	person from Canada	0.0000	0.0000
DeepSeek-Coder-V2	person from Australia	0.0833	0.0870
DeepSeek-Coder-V2	person from New Zealand	0.0000	-0.1304
DeepSeek-Coder-V2	person from South Africa	0.0000	0.0000

Table 4: Explicit stereotypes for DeepSeek models elicited through persona prompting.

B Implicit and Explicit Language Stereotypes

Implicit and explicit language stereotypes for each model are presented in Tables 5 and 6.

Model	Persona	Economic Stereotype	Social Stereotype
Gemini-2.0-flash-lite	Italian speaking person	-0.033	-0.136
Gemini-2.0-flash-lite	German speaking person	0.133	-0.180
Gemini-2.0-flash-lite	French speaking person	0.117	-0.120
Gemini-2.0-flash-lite	Polish speaking person	-0.033	-0.103
Gemini-2.0-flash-lite	Czech speaking person	0.150	-0.136
Gemini-2.0-flash-lite	Spanish speaking person	0.067	-0.136
Gemini-2.0-flash-lite	Questions Italian	-0.117	-0.188
Gemini-2.0-flash-lite	Questions German	-0.200	-0.334
Gemini-2.0-flash-lite	Questions French	0.133	-0.116
Gemini-2.0-flash-lite	Questions Polish	0.233	-0.110
Gemini-2.0-flash-lite	Questions Czech	0.355	0.364
Gemini-2.0-flash-lite	Questions Spanish	0.233	0.019
Gemini-2.5-flash	Italian speaking person	-0.200	-0.143
Gemini-2.5-flash	German speaking person	-0.117	-0.214
Gemini-2.5-flash	French speaking person	-0.300	-0.214
Gemini-2.5-flash	Polish speaking person	-0.100	-0.071
Gemini-2.5-flash	Czech speaking person	-0.200	-0.143
Gemini-2.5-flash	Spanish speaking person	-0.200	0.067
Gemini-2.5-flash	Questions Italian	-0.383	-0.186
Gemini-2.5-flash	Questions German	-0.283	-0.186
Gemini-2.5-flash	Questions French	-0.283	-0.186
Gemini-2.5-flash	Questions Polish	-0.017	0.174
Gemini-2.5-flash	Questions Czech	-0.117	0.138
Gemini-2.5-flash	Questions Spanish	-0.300	-0.143
Gemini-2.0-flash	Italian speaking person	-0.008	0.018
Gemini-2.0-flash	German speaking person	0.008	-0.001
Gemini-2.0-flash	French speaking person	-0.008	-0.027
Gemini-2.0-flash	Polish speaking person	0.008	-0.047
Gemini-2.0-flash	Czech speaking person	-0.008	-0.099
Gemini-2.0-flash	Spanish speaking person	-0.108	-0.034
Gemini-2.0-flash	Questions Italian	0.092	-0.150
Gemini-2.0-flash	Questions German	-0.008	0.011
Gemini-2.0-flash	Questions French	0.242	0.220
Gemini-2.0-flash	Questions Polish	0.075	0.199
Gemini-2.0-flash	Questions Czech	-0.192	0.178
Gemini-2.0-flash	Questions Spanish	0.175	0.059
GPT-4.1-mini	Italian speaking person	0.092	0.644
GPT-4.1-mini	German speaking person	0.258	0.430
GPT-4.1-mini	French speaking person	0.175	0.335
GPT-4.1-mini	Polish speaking person	0.008	0.388
GPT-4.1-mini	Czech speaking person	0.258	0.347
GPT-4.1-mini	Spanish speaking person	0.092	0.263
GPT-4.1-mini	Questions Italian	0.158	0.644
GPT-4.1-mini	Questions German	-0.008	0.585
GPT-4.1-mini	Questions French	0.158	0.472
GPT-4.1-mini	Questions Polish	0.007	0.421
GPT-4.1-mini	Questions Czech	0.492	0.567
GPT-4.1-mini	Questions Spanish	0.175	0.335

Table 5: Implicit and explicit language stereotypes for GPT and Gemini models.

Model	Persona	Economic Stereotype	Social Stereotype
Llama-4	Italian speaking person	0.383	0.071
Llama-4	German speaking person	0.283	0.115
Llama-4	French speaking person	0.183	0.000
Llama-4	Polish speaking person	0.383	0.143
Llama-4	Czech speaking person	0.467	0.071
Llama-4	Spanish speaking person	0.367	0.115
Llama-4	Questions Italian	0.267	-0.071
Llama-4	Questions German	0.267	0.087
Llama-4	Questions French	0.367	-0.087
Llama-4	Questions Polish	0.267	-0.043
Llama-4	Questions Czech	0.267	0.087
Llama-4	Questions Spanish	0.283	0.071
Llama-3.3	Italian speaking person	0.083	0.043
Llama-3.3	German speaking person	0.083	0.043
Llama-3.3	French speaking person	0.000	0.000
Llama-3.3	Polish speaking person	0.083	0.043
Llama-3.3	Czech speaking person	0.083	0.043
Llama-3.3	Spanish speaking person	0.083	0.043
Llama-3.3	Questions Italian	0.000	0.043
Llama-3.3	Questions German	0.083	-0.028
Llama-3.3	Questions French	0.083	0.043
Llama-3.3	Questions Polish	0.083	0.130
Llama-3.3	Questions Czech	0.250	0.115
Llama-3.3	Questions Spanish	0.000	0.130
DeepSeek-R1	Italian speaking person	0.000	0.119
DeepSeek-R1	German speaking person	0.000	0.000
DeepSeek-R1	French speaking person	0.017	0.119
DeepSeek-R1	Polish speaking person	0.100	0.143
DeepSeek-R1	Czech speaking person	0.017	0.190
DeepSeek-R1	Spanish speaking person	-0.083	-0.043
DeepSeek-R1	Questions Italian	-0.083	-0.043
DeepSeek-R1	Questions German	0.100	-0.043
DeepSeek-R1	Questions French	0.083	-0.071
DeepSeek-R1	Questions Polish	0.100	0.164
DeepSeek-R1	Questions Czech	0.200	0.158
DeepSeek-R1	Questions Spanish	0.200	0.164
DeepSeek-Coder-V2	Italian speaking person	0.350	0.043
DeepSeek-Coder-V2	German speaking person	0.383	0.000
DeepSeek-Coder-V2	French speaking person	0.183	0.056
DeepSeek-Coder-V2	Polish speaking person	0.367	-0.028
DeepSeek-Coder-V2	Czech speaking person	0.350	-0.043
DeepSeek-Coder-V2	Spanish speaking person	0.350	0.000
DeepSeek-Coder-V2	Questions Italian	0.483	-0.130
DeepSeek-Coder-V2	Questions German	0.467	-0.043
DeepSeek-Coder-V2	Questions French	0.283	-0.143
DeepSeek-Coder-V2	Questions Polish	0.283	0.190
DeepSeek-Coder-V2	Questions Czech	0.267	0.258
DeepSeek-Coder-V2	Questions Spanish	0.083	-0.202

Table 6: Implicit and explicit language stereotypes for Llama and DeepSeek models.