Through the Looking Glass of Multilingual AI: Contrasting Language- and Name Script-Dependent Ethnic Hierarchies in GPT and DeepSeek

Anonymous Author(s)

Abstract

Large language models (LLMs) are increasingly used as evaluative tools across languages, yet bias research remains overwhelmingly Anglocentric as much work is conducted primarily in English with Latin-script names. It remains unclear whether bias patterns generalize across linguistic contexts. We investigate this question and find that prompt language and writing script alter the hierarchical ordering of ethnic groups and reveal model behaviors that monolingual evaluations fail to detect. We conduct a large-scale study comprising 900,000 model responses based on 45,000 names in 9 ethnicities, systemetically varying models (GPT, DeepSeek), prompt languages (English, Chinese, Thai), writing scripts (Latin, Chinese, Thai), and evaluation domains (competence and warmth dimensions). We also evaluate the stability of ethnic rankings across linguistic conditions and models. Results reveal that ethnic name biases differ substantially between Western-centric and Sinocentric models. DeepSeek maintains nearly identical rankings across languages and name scripts in math competence judgment, with Chinese always at the top and Russian second and the bottom three being White, Hispanic, and Black. GPT, on the contrary, exhibits language- and script-dependent reordering, with a high degree of ranking stability in two distinct clusters: when ethnic names are written in Latin (romanized) scripts and when names are transliterated into local non-English scripts matching non-English prompt languages. Second, warmth dimension evaluations are less stable: although Iranian and Russian names tend be biased against, GPT shows substantial cross-language rank changes. DeepSeek's rankings also shift substantially under native-script conditions.

More broadly, our study demonstrates that multilingual bias cannot be characterized by single-language, single-writing system audits. An organization auditing for bias only in English with romanized names may observe some levels of ethnic variation, yet the same models, deployed in another language with transliterated names, can produce substantially different or even inversely correlated ethnic hierarchies. For multilingual users, code-switching between languages may unknowingly toggle between different bias regimes, with potential consequences for any application where LLMs assess human competence and character. Fairness evaluations for multilingual LLMs should therefore test across deployment languages, writing scripts, evaluation domains, and models to capture the full range of bias these systems carry.

ACM Reference Format:
Anonymous Author(s). 2018. Through the Looking Glass of Multilingual AI: Contrasting Language- and Name Script-Dependent Ethnic Hierarchies in GPT and DeepSeek. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

> "Mtee Tet ain't American lmfao."
> "What is it that makes you say [Karn Chutinan] isn't American?" (In response to comments regarding the USA International Math Olympiad team members' ethnic last names)
> "People who want to assimilate to America don't name their kids 'Savithri.'"

These real social media reactions to children's names show how quickly personal names become bases of exclusion and ethnocentric judgment. In early 2026, U.S. politician and former presidential candidate Vivek Ramaswamy announced his daughter's name *Savithri*, prompting backlash that framed the name as "not American enough." Rather than being treated as neutral identifiers, names were read as signals of belonging, authenticity, and assimilation, often with social penalties for those deemed outside the cultural mainstream. These reactions illustrate a broader concern in our multicultural societies: ethnic names do not merely reflect identity but they also often reveal underlying cultural prejudices about who is perceived as legitimate, competent, trustworthy, or truly belonging.

Increasingly, LLMs are used to assist or substitute for human evaluators in high-stakes domains such as education, healthcare, and hiring, which shape life chances and access to resources. Growing evidence shows that these models carry covert sociocultural biases that shape how people are evaluated, even when protected characteristics are not explicitly mentioned. Yet most existing work has focused on broad racial categories within English-dominant settings, leaving a critical gap in our understanding of how bias operates across languages, writing systems, and more granular identity signals.

This paper investigates ethnic names as a systematic, underexplored lens for uncovering multilingual AI bias. Names are not merely labels; they signal rich social information about ethnicity, religion, gender, migration history, and class. They also activate social heuristics such as stereotypes and ethnocentrism, the tendency to favor one's perceived in-group while stereotyping or devaluing out-groups. Decades of social science research show that ethnocentrism shapes how people interpret information, assess competence,

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

and allocate trust. However, whether LLMs exhibit analogous patterns of in-group favoritism across languages and cultural contexts remains largely unknown. Furthermore, although it has been found that LLMs rank individuals hierarchically based on their name ethnicity [20], the experiment was conducted in English and on LLMs originated in the United States. By systematically varying first and last names across 9 ethnicities, 3 languages (English, Chinese, Thai), and two universal, orthogonal human dimensions (competence and warmth), this work examines whether LLMs privilege culturally proximate names to the prompt languages, reproduce implicit hierarchies among ethnic groups, or treat transliterated identities differently. Our experiments, analyzing 900,000 LLM responses, reveal that LLM bias operates through multiple interacting mechanisms rather than a single axis of discrimination. We find stark contrasts between GPT's and DeepSeek's biases and hierarchies. Overall, we show significant ethnic name bias in multilingual LLMs in which ranking structures and their stability depend on linguistic context. Multilingual context reorganizes bias ranking hierarchies, and models differ in how stable those structures are. These findings demonstrate that fairness evaluations based on single-language testing does not hold for multilingual LLMs, and LLM evaluations should systematically vary prompt language, writing script, evaluation domain, and model choice to understand bias exhibited by multilingual AI systems in an era of global AI deployment.

2 Background

2.1 Stereotype and Bias

Humans naturally categorize both themselves and others into social groups, often along boundaries such as ethnicity. Once formed, categories become the basis of prejudgment [2]. Although such categorization may once have served survival purposes, modern social categories are embedded with beliefs that can give rise to bias and stereotypes, which are biased thoughts and beliefs about a person or a social group due to beliefs that the category describes them [8, 9]. The term originally referred to a printer's metal plate that could hold an entire page of print, allowing printers to produce identical copies of a page. Walter Lippmann described stereotypes as "the pictures in our heads" that present members of a group to have the same attributes [8]. While stereotypes may originate from observations, they are often exaggerated and restrictive, shaping expectations and limiting opportunities [17]. Relatedly, prejudice refers to evaluative attitudes or biases against people based on their group membership [10]. While stereotypes are thoughts and beliefs, prejudice are evaluative, the attitudes about social groups. Historically, prejudice was frequently overt, with laws and social norms openly disadvantaging minoritized groups. For example, research in the United States has documented persistent stereotypes portraying Black individuals as less competent, and less trustworthy [13], while early studies revealed explicit negative views toward Jewish people [15].

Much research has examined racial and ethnic prejudice, particularly systems that historically advantaged White populations and marginalized people of color. In particular, the anti-Black prejudice is shown to be especially pervasive in the United States [17]. Although overt bias has declined, prejudice often persists in subtle and implicit forms that continue to influence behavior and resource allocation. Bias can also change over time, with newer immigrant groups frequently becoming targets of hostility and dehumanizing stereotypes [8, 9]. While numerous stereotypes exist, Fiske et al. [10] propose that they can be organized along two core dimensions: warmth and competence, which reflect judgments about others' intentions and abilities. Groups perceived as high in both warmth and competence elicit admiration, those low on both evoke contempt, high competence but low warmth produces envy, and high warmth but low competence leads to paternalistic attitudes.

2.1.1 Ethnic Biases in LLMs. Large language models (LLMs) have rapidly become ubiquitous and active participants in evaluative workflows, often operating alongside human decision-makers. In domains such as education, healthcare, hiring, finance, and governance, LLM-based systems increasingly assist human evaluators in judging competence and credibility of individuals, shaping their life chances, social mobility, and access to resources. In this hybrid human–AI landscape, fairness, accountability, and social harm are central challenges for designing fair AI evaluation agents. Growing evidence suggests that they can reproduce, transform, and amplify deeply rooted social inequalities in subtle ways that are difficult to detect through surface-level evaluations. Recent work has demonstrated that LLMs encode covert forms of prejudice that emerge in context-dependent evaluation [16]. For example, LLMs exhibit anti-Muslim bias [1], treat text differently based on dialectal cues such as African American English [13], discriminate between otherwise identical resumes when names signal race or gender, and alter medical recommendations based solely on sociodemographic labels such as race, housing status, or LGBTQIA+ identity. Hofmann et al. [13] found that LLMs covertly exhibit highly negative archaic stereotypes of speakers of African American English, and the associations are much more harmful than what the models overtly claim. These findings indicate that LLMs carry sociocultural priors that shape how people are evaluated, categorized, and treated, even when protected characteristics are not explicitly mentioned. Yet most existing work has focused on broad racial categories within English-dominant settings, leaving a critical gap in our understanding of how bias operates across cultures, languages, and more granular ethnic identity signals.

2.2 Names

Several recent works have studied name biases in language models [4, 14, 18, 21, 23–25]. An et al. [3] studied 300 White, Black, and Hispanic first names and found that LLMs tend to favor White applicants in hiring decisions, while Hispanic names receive the least favorable treatment.

Sakunkoo and Sakunkoo [20] analyzed 45,000 name variations across five ethnicities and found that LLMs construct status hierarchies based on names signaling race and gender. Rather than exhibiting uniform White favoritism, they find that East Asian names are often ranked highest in perceived academic competence while Southeast Asian names are consistently ranked lowest, complicating both simplistic racial models of bias and the monolithic "model minority" stereotype. While this work powerfully documents name-based stratification, it is conducted primarily within English-language prompts, leaving open questions about how such hierarchies shift across languages, scripts, and cultural contexts.

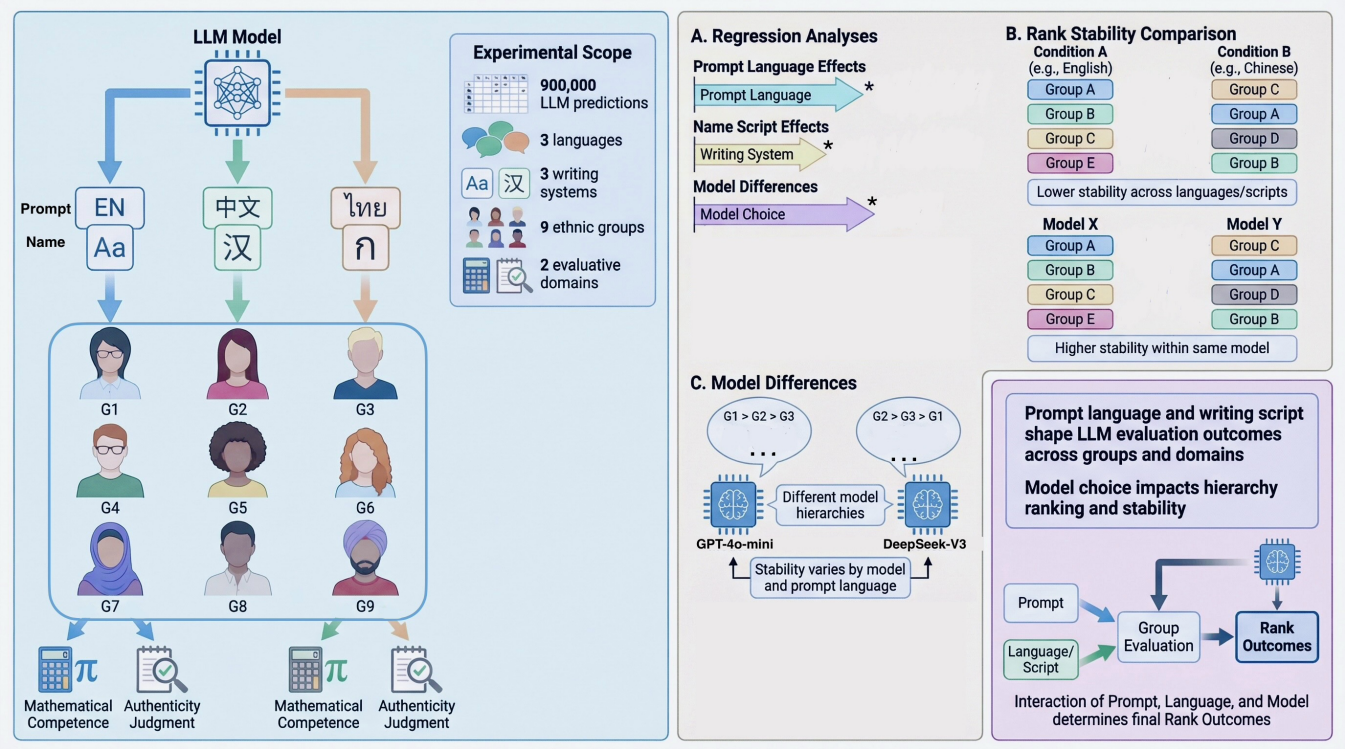

Figure 1: Investigating and Contrasting AI Ethnic Name Prejudice in GPT and DeepSeek

Building on these prior insights, this paper provides large-scale empirical evidence comparing and contrasting an American-based LLM and a Chinese-based LLM in a multilingual setting and examining whether and how ethnic name biases emerge when the prompt language of interaction and the linguistic scripts of the names change.

We address the following research questions:

Q1: What bias and hierarchical patterns are evident in ChatGPT's and DeepSeek's responses?

Q2: How do these patterns differ across the two LLMs (American vs Chinese), prompt languages, linguistic name scripts, and subject?

3 Experiment Setup

Figure 1 provides an overview of our experimental design. We investigate how prompt language, writing script, and model choice jointly shape ethnic bias in LLM evaluations. Our design varies four factors: model (GPT-4o-mini, DeepSeek-V3), prompt language (English, Chinese, Thai), writing script (Latin, native), and evaluation domain (competence, warmth dimensions). We test language–script combinations likely to occur in real deployment: English prompts with Latin-script names, Chinese prompts with both Latin-script and Chinese-script names, and Thai prompts with both Latin-script and Thai-script names. We select two models developed by organizations in different cultural and regulatory contexts, GPT-4o-mini (OpenAI, US-based) and DeepSeek-V3 (DeepSeek, China-based), to test whether bias patterns reflect model-specific distinct institutional, cultural, and alignment contexts in which they were developed rather than universal stereotypes.

Name Data We evaluate nine ethnic groups: Chinese, Russian, Jewish, Indian, Iranian, Thai, White (European-American), Hispanic, and Black (African-American). These groups were selected to span geographic, cultural, geopolitical, and ethnic categories and to include groups for which distinct native-script name forms exist (Chinese and Thai).

For each ethnic group, we compile prototypical first and last names that are associated with the target ethnicity by natives of the relevant ethnic background. The names are evenly distributed between female and male. With 100 first names and 50 last names for each ethnicity, we have 5,000 unique name variations, thus 45,000 unique name variations in total. (Name selection details are available in Appendix A.) Names are presented in two script conditions: Latin script (romanized forms used across all prompt languages) and native prompt-language script.

Prompts Mathematical competence serves as a competence-dimension measure in the Stereotype Content Model [10], anchored to an objective scale and therefore expected to be relatively resistant to cross-linguistic semantic drift. Authenticity is a warmth-dimension construct, a character judgment central to hiring decisions, university admissions, and appeal [12, 19]. All prompts follow a template that requests a numerical prediction for a named individual. Each prompt specifies a named individual (with first and last names drawn from the ethnic name pool in the appropriate

script) and an evaluation task (competition math score prediction and authenticity rating). The name-substitution methodology for detecting bias is well established [3, 5, 6, 11], and our prompts contain only the applicant's first and last names with no additional biographical details; education, experience, or other demographic information are deliberately omitted so that any systematic variation in model predictions can be attributed solely to the name's ethnic signal rather than confounding applicant characteristics [22]. Numerical scores are then extracted from model responses.

Regression Analyses For each of the 20 condition combinations, we employ ordinary least squares (OLS) regression to analyze how the LLM assigns math competence and authenticity scores based on ethnicity and gender, through student first and last names. We use bootstrap resampling (1,000 replications) to estimate coefficient variability and ensure robust inferences. We then rank-order the nine ethnic groups by their coefficients within each condition to produce an ethnic bias hierarchy.

Rank Stability Comparison To assess whether bias hierarchies are stable across conditions, we compute pairwise Spearman rank correlations between all condition pairs within and across models. This evaluates whether the relative ordering of groups is preserved, a more informative measure than comparing raw score differences, which vary in scale across domains and models.

Cross-Model Comparison Finally, we directly compare the two models' ethnic hierarchies under matched conditions to identify where they converge and diverge. This reveals whether bias patterns are shared properties of LLMs in general or reflect model-specific cultural origins and whether there are differences between standard English-language evaluation and multilingual testing.

LLM Models We conduct our experiments on ethnic name biases using GPT4o-mini and DeepSeek-V3 to test whether they differ in ways that align with their companies' cultural origins (Western-centric and Sinocentric, respectively).

4 Results and Discussion

4.1 Predicted Math Competence

A consistent finding across both models is that mathematical bias hierarchies are stable within models, but the degree of stability and the hierarchies differ remarkably between models. DeepSeek exhibits virtually no script and prompt language effect. All ten pairwise correlations among its five math conditions fall at or above $\rho = 0.90$, with Chinese names maintaining rank 1 and Russian names rank 2 across every condition regardless of prompt language or script. The bottom three are also consistently White, Hispanic, and then Black names. The near-complete invariance of DeepSeek's math hierarchy implies that whatever stereotypic associations drive mathematical competence prejudice in DeepSeek, they are deeply embedded and resistant to surface-level linguistic manipulation.

GPT produces math rankings that cluster tightly by script type. The three Latin-script conditions (English prompt-Latin name script, Chinese prompt-Latin name script, Thai prompt-Latin name script) intercorrelate at a mean $\rho = 0.94$, and the two native-script conditions (Chinese prompt-Chinese name script, Thai prompt-Thai name script) correlate with each other at $\rho = 0.97$. However, cross-cluster correlations drop to a mean $\rho = 0.66$, a substantial decline indicating that switching from Latin to native name script reorganizes the hierarchy. The most striking shift in GPT is positional: Chinese names rank 1st or 2nd across all Latin-script conditions, but White names, which are ranked 3rd in Latin-script conditions, rise to 1st in both native-script conditions. Also, while the Thai prompt, Thai name script condition places Chinese 6th in the hierarchy, switching the name script to Latin script pushes Chinese students to 1st in the hierarchy. This is, to our knowledge, the first empirical demonstration that a leading multilingual LLM can produce inversely correlated ethnic bias hierarchies under matched conditions. In stark contrast to DeepSeek's hierarchy, Iranian and Thai names consistently rank at the bottom in GPT's mathematical hierarchy. This pattern, which we term script-gated bias, suggests that native-script contexts may activate a different evaluative frame in GPT, as prompt language-native script matching elevates Western-associated names to the top of the competence hierarchy.

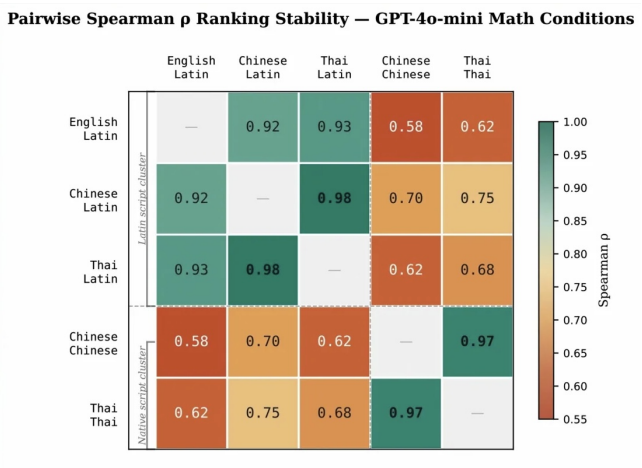

Figure 2: Pairwise Rank Order Correlations of GPT math competition conditions. Latin name script conditions form a tight cluster, while native name script conditions cluster separately.

These contrasting patterns reveal that hierarchical ethnic prejudice differs across models, and within-model stability is not a general property of LLMs but a model-specific characteristic. The same experimental manipulation such as switching from Latin to native script produces negligible effects in one model (DeepSeek) but substantial hierarchy reorganization in another (GPT).

4.2 Authenticity

Authenticity hierarchies are far less stable than mathematical ones in both models. GPT's authenticity rankings show weak within-model consistency. The English condition and Thai-Thai condition show no correlation at all, indicating that the ethnic hierarchy GPT produces for authenticity in English bears no structural relationship to the one it produces in the Thai script. Thai names, ranked 9th (last) in the English condition, rises to 1st in the Thai condition, a full inversion rather than a slight shift. Furthermore, Indian names rank 1st in the English condition, but White names rank 1st in the

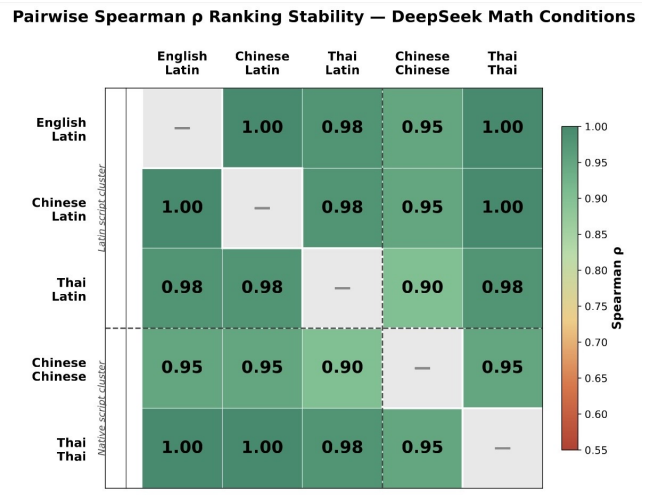

Figure 3: Pairwise Rank Order Correlations of DeepSeek math competition conditions show high stability of ethnic rankings across linguistic conditions.

Chinese linguistic conditions. For GPT, Iranians and Russians are consistently ranked in the bottom three in terms of perceived authenticity. DeepSeek's authenticity rankings are more stable than

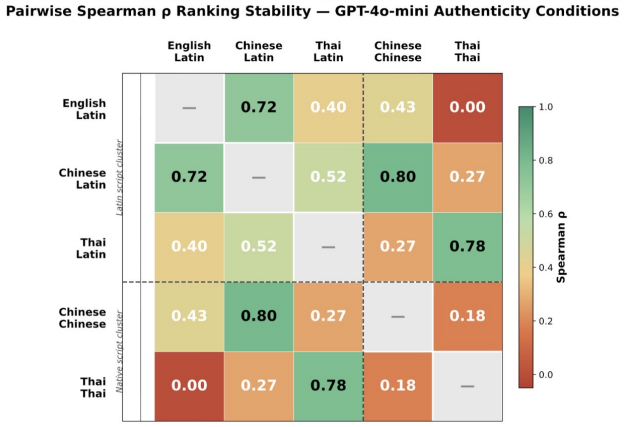

Figure 4: Pairwise Rank Order Correlations of GPT authenticity conditions

GPT but still far less stable than its own math rankings. Iranian and Russian names, again, are ranked in the bottom three in all conditions. While Chinese individuals are ranked 4th in the English condition, they are consistently in the top two in non-English linguistic conditions.

The strongest ethnocentrism of any condition emerges in DeepSeek's Thai-Thai authenticity condition, where Thai names rise to 1st and deviate +6.18 points above the group mean, showing that the Thai prompt/Thai name script condition results in the model giving Thai people much higher authenticity rating. This in-group favoritism is invisible under the Thai prompt, Latin-script testing, where Thai

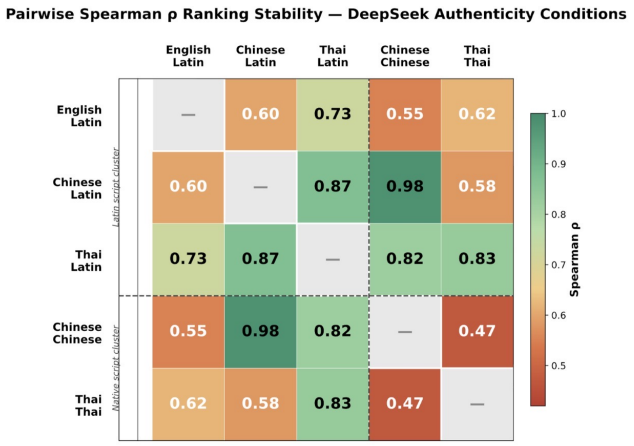

Figure 5: Pairwise Rank Order Correlations of DeepSeek authenticity conditions

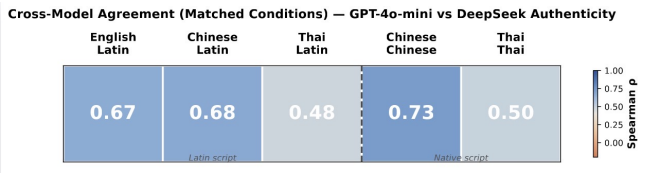

Figure 6: Spearman ρ Correlations of GPT and DeepSeek authenticity conditions

names rank 5th, revealing it as a purely script-gated effect. The results hence show a script-gated bias pattern that is invisible when testing with romanized names (Latin script) alone.

5 Conclusion

Across all conditions and human dimensions, a persistent pattern emerges: each LLM model carries ethnic biases that reflect, at least partially, the cultural context of its development. The rankings are statistically significant, and certain ethnicities consistently cluster at the top or bottom. However, the specific shape of those biases is malleable as it shifts based on prompt languages, writing scripts, and subjects. GPT ranks White names 1st or near the top across all five authenticity conditions regardless of prompt language or script, revealing a Western-centric default that persists even when the model is used in non-English languages. DeepSeek consistently places Chinese names at or near the top across both domains and all script conditions. In math, this Sinocentric bias of DeepSeek is absolute: Chinese holds rank 1 in every condition. In authenticity, it is strong but modulated by script. We have shown that ethnic bias in LLMs is not a fixed property but a dynamic system shaped by the interaction of prompt language, writing script, evaluation domain, and model provenance. The bias patterns shown in our study such as hierarchy inversions in authenticity and script-gated ethnocentrism are invisible under romanized testing and emerge only when evaluation moves beyond English. Many multilingual users may unknowingly traverse different bias regimes. A bilingual

user who switches between English and another language when interacting with the same model may receive evaluations governed by different ethnic hierarchies. Code-switching, a natural behavior for the global majority of internet users, also functions as an inadvertent bias toggle. As LLMs are increasingly deployed as evaluative tools for the global majority who do not operate solely in English and romanized name scripts, fairness research must follow them there.

Limitations

This study includes only nine ethnicities, out of numerous other ethnic identities. The study's decision does not suggest that other ethnicities are not important. We also acknowledge potential limitations in our name dataset, as discussed in Appendix A. Additionally, names can reflect other attributes such as religion and age. Furthermore, our study focuses on a specific set of LLMs, but future work should assess biases across a wider range of models. Exploring LLMs in more than two non-English languages would also uncover distinct patterns of bias and social hierarchies that are not captured in this study.

Our study uses a minimal-context design to isolate how LLMs respond to names alone, without additional context. This approach aims to detect bias and reveal whether an LLM's response is influenced by the mere differences in names associated with race and gender as it makes biased predictions with different names even before any substantive input is given. However, we acknowledge that this design does not illustrate how such biases might affect individuals when more contextual profiles are involved. Future work will build on this foundation by including more relevant inputs [7].

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306. doi:10.1145/3461702.3462624

[2] Gordon W. Allport. 1954. *The Nature of Prejudice*. "Addison-Wesley".

[3] Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 386–397. doi:10.18653/v1/2024.acl-short.37

[4] Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-Ended Discovery of Social Biases in Social Commonsense Reasoning Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 1573–1596. doi:10.18653/v1/2023.eacl-main.116

[5] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (2004), 991–1013. doi:10.1257/0002828042002561

[6] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. doi:10.1126/science.aal4230

[7] Kathryn Campbell-Kibler. 2008. I'll be the judge of that: Diversity in social perceptions of (ING). *Language in Society* 37, 5 (2008), 637–659. doi:10.1017/S0047404508080974

[8] Jennifer L. Eberhardt. 2019. *Biased: Uncovering the Hidden Prejudice That Shapes What We See, Think, and Do*. Viking. https://www.penguinrandomhouse.com/books/557462/biased-by-jennifer-l-eberhardt-phd/

[9] Susan T Fiske. [n. d.]. Prejudice, discrimination, and stereotyping. http://noba.to/jfkx7nrd. Accessed: 2026-2-18.

[10] Susan T Fiske, Amy J C Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 2 (Feb. 2007), 77–83.

[11] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74, 6 (1998), 1464–1480. doi:10.1037/0022-3514.74.6.1464

[12] Anna Luca Heimann and Annika Schmitz-Wilhelmy. 2024. Observing interviewees' inner self: How authenticity cues in job interviews relate to interview and job performance. *J. Bus. Psychol.* (May 2024).

[13] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, 8028 (Sept. 2024), 147–154.

[14] Sullam Jeoung, Jana Diesner, and Halil Kilicoglu. 2023. Examining the Causal Impact of First Names on Language Models: The Case of Social Commonsense Reasoning. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 61–72. doi:10.18653/v1/2023.trustnlp-1.7

[15] D Katz and K Braly. 1933. Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology* 28, 3 (1933), 280–290.

[16] F W Kerche, M Zook, and M Graham. 2026. The silicon gaze: A typology of biases and inequality in LLMs through the lens of place. *Platforms & Society* 3 (2026).

[17] Mary E Kite, Bernard E Whitley, Jr, and Lisa S Wagner. 2022. *Psychology of prejudice and discrimination* (4 ed.). Routledge, London, England.

[18] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5267–5275. doi:10.18653/v1/D19-1530

[19] Kieran O'Connor, Glenn R Carroll, and Balázs Kovács. 2017. Disambiguating authenticity: Interpretations of value and appeal. *PLoS One* 12, 6 (June 2017), e0179187.

[20] Annabella Sakunkoo and Jonathan Sakunkoo. 2025. Name of Thrones: How Do LLMs Rank Student Names in Status Hierarchies Based on Race and Gender?. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, Ekaterina Kochmar, Bashar Alhafni, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Vienna, Austria, 697–707. doi:10.18653/v1/2025.bea-1.50

[21] Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "You are grounded!": Latent Name Artifacts in Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6850–6861. doi:10.18653/v1/2020.emnlp-main.556

[22] Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023. Are Emily and Greg Still More Employable than Lakisha and Jamal? Investigating Algorithmic Hiring Bias in the Era of ChatGPT. *CoRR* abs/2310.05135 (2023). arXiv:2310.05135 doi:10.48550/ARXIV.2310.05135

[23] Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*.

[24] Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and Mitigating Name Biases in Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2576–2590. doi:10.18653/v1/2022.acl-long.184

[25] Robert Wolfe and Aylin Caliskan. 2021. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 518–532. doi:10.18653/v1/2021.emnlp-main.41

A Name Data

We compiled first and last names from several primary sources: national delegates of academic and music competitions and the most frequently occurring names in population databases. To construct our dataset, we randomly sampled names from these sources and ensured balanced representation across different origins. To validate the accuracy of name classification, we had native speakers

from each cultural background verify that the selected names are characteristic of their respective origins and gender. Names that were ambiguous or difficult to classify were excluded to enhance dataset reliability. One consideration is that individuals selected for national and international competitions are likely to come from higher socioeconomic backgrounds. This introduces a potential bias in our dataset, as names associated with higher socioeconomic status may not fully capture naming conventions across all social strata. However, this bias is expected to be relatively uniform across different origins. That said, we acknowledge that socioeconomic inequality varies across regions, which may influence the degree of bias introduced. This remains a limitation and an avenue for further research. Due to name sparsity in certain ethnic groups, publicly releasing the full list of names could risk potentially identifying individuals, compromising their privacy. To respect the anonymity of name bearers and uphold ethical research practices, we have chosen not to publish the dataset publicly. However, researchers interested in the name lists may contact the authors for access under appropriate research agreements.

B Sample Ethnic Hierarchies

We include samples of ethnic hierarchies in this paper, and all regression analyses yield statistically significant results. The full regression results will be presented and discussed at the workshop.

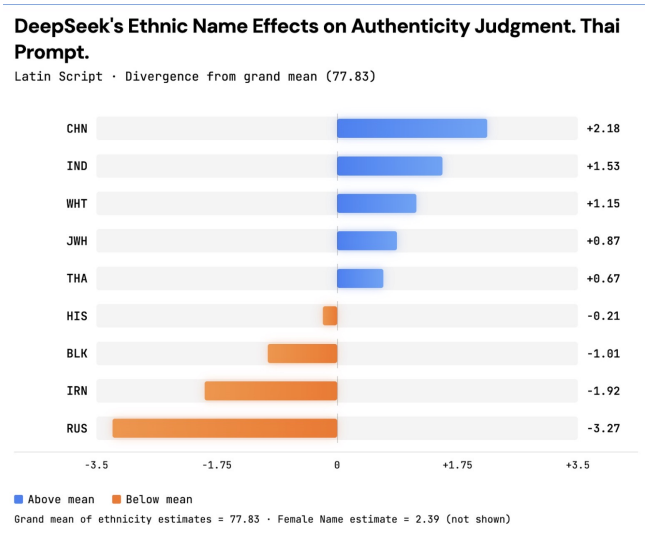

Figure 7

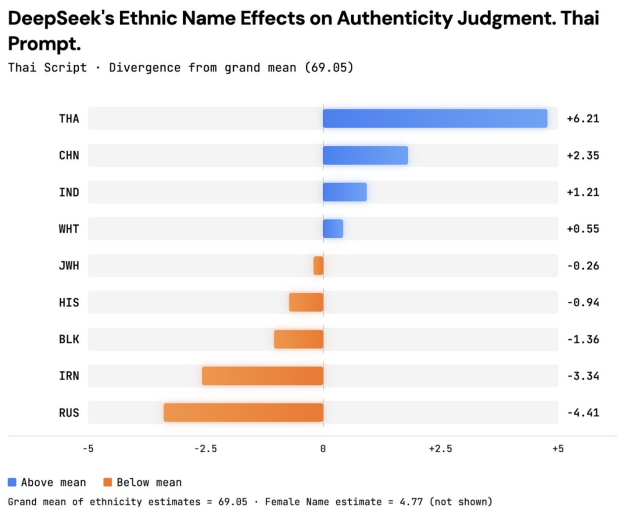

Figure 8: Fig. 7-8 show ethnic hierarchies; same Thai prompt; Latin vs. Thai name script (Thai rises to No.1)

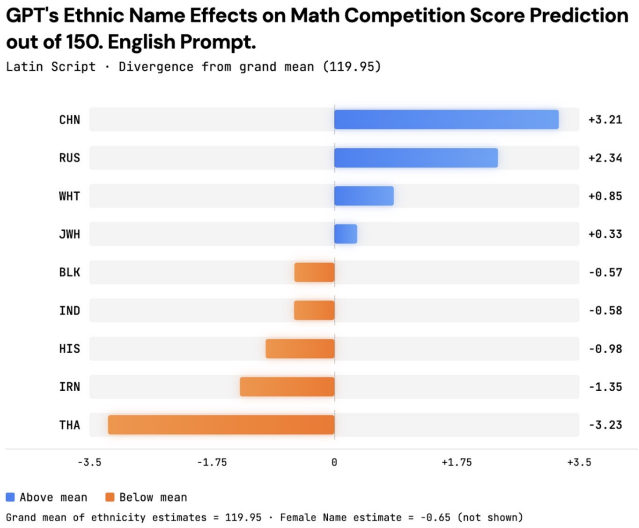

Figure 9

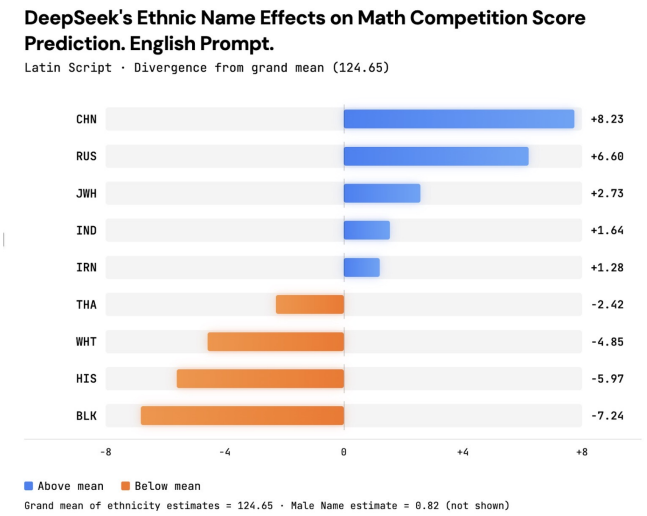

Figure 10: Fig. 9-10 show ethnic rankings of math competition predictions by GPT vs DeepSeek