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

Annabella Sakunkoo
Stanford University OHS
apianist@ohs.stanford.edu

Jonathan Sakunkoo
University of Oxford
jonathan.sakunkoo@cs.ox.ac.uk

Abstract

Large language models (LLMs) are increasingly used as evaluative tools across languages, yet bias research remains overwhelmingly Anglocentric, with most studies conducted in English using Latin-script names. It remains unclear whether bias patterns generalize across linguistic contexts. We investigate this question and introduce the stereotype perceptual map, a framework for analyzing how ethnic groups are positioned along evaluative dimensions. We find that prompt language and writing script alter the hierarchical ordering of ethnic groups and reveal model behaviors that monolingual evaluations fail to detect. Using 900,000 model responses over 45,000 name variations spanning 9 ethnicities, we evaluate model behavior across prompt languages (English, Chinese, Thai), writing scripts (Latin, Chinese, Thai), evaluative domains (competence, warmth), and models (GPT, DeepSeek). We find that ethnic bias hierarchies are jointly shaped by local linguistic context and model origin and differ substantially between Western-centric and Sinocentric models. DeepSeek exhibits highly stable rankings across conditions in math competence judgments, consistently placing Chinese at the top, followed by Russian, and White, Hispanic, and Black names at the bottom. GPT, by contrast, shows strong script-dependent reordering: Latin-script conditions form one stable cluster, while native-script conditions form another, with substantially lower cross-cluster correlations. We term this script-gated bias: transliterating the same names into a non-Latin script can activate a different evaluative frame and produce rankings that are sometimes inversely correlated with Latin-script results. Warmth evaluations are less stable than competence in both models.

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 unknowingly toggles 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 model origins to capture the full range of potentially harmful bias these systems carry.

1 Introduction

> "Mtee Tet ain't American lmfao."
>
> "What is it that makes you say [K... Chutinan] isn't American?" (In response to comments regarding the USA International Math Olympiad team members' ethnic 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 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 deservingness, 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.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1103–1114

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. It remains unclear whether these bias patterns persist when the same names are evaluated across different languages and writing systems.

This paper investigates ethnic names as a systematic, underexplored lens for uncovering multilingual AI bias. Names 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, 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 (Sakunkoo and Sakunkoo, 2025), 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, produce 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. Multilingual context reorganizes ranking hierarchies, and models differ in how stable those biases are. These findings demonstrate that fairness evaluations based on single-language testing do not hold for multilingual LLMs, and LLM evaluations should systematically vary prompt language, writing script, evaluation domain, and model choice to understand harmful 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 (Allport, 1954). Although such categorization may once have served survival purposes, modern social categories are filled with stereotypes, which are biased thoughts and beliefs about a person or social group due to their social category (Fiske, 2026; Eberhardt, 2019). 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 (Eberhardt, 2019). While stereotypes may originate from observations, they are often exaggerated and restrictive, shaping expectations and limiting opportunities (Kite et al., 2022). Relatedly, prejudice refers to evaluative attitudes or biases against people based on their group membership (Fiske et al., 2007).

Historically, stereotypes and prejudice were frequently overt, openly disadvantaging minoritized groups. For example, research in the United States has documented persistent stereotypes portraying Black individuals as less competent and less trustworthy (Hofmann et al., 2024), while early studies revealed explicit negative views toward Jewish people (Katz and Braly, 1933). In particular, anti-Black prejudice was especially pervasive in the United States (Kite et al., 2022). Although overt bias has declined over time, it 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 (Eberhardt, 2019).

Although numerous stereotypes exist, they can be organized along two core dimensions (Fiske et al., 2007): competence and warmth, which reflect judgments about others' abilities and intentions. Groups perceived as high in both warmth and competence elicit admiration, those low on both evoke contempt, high competence but low

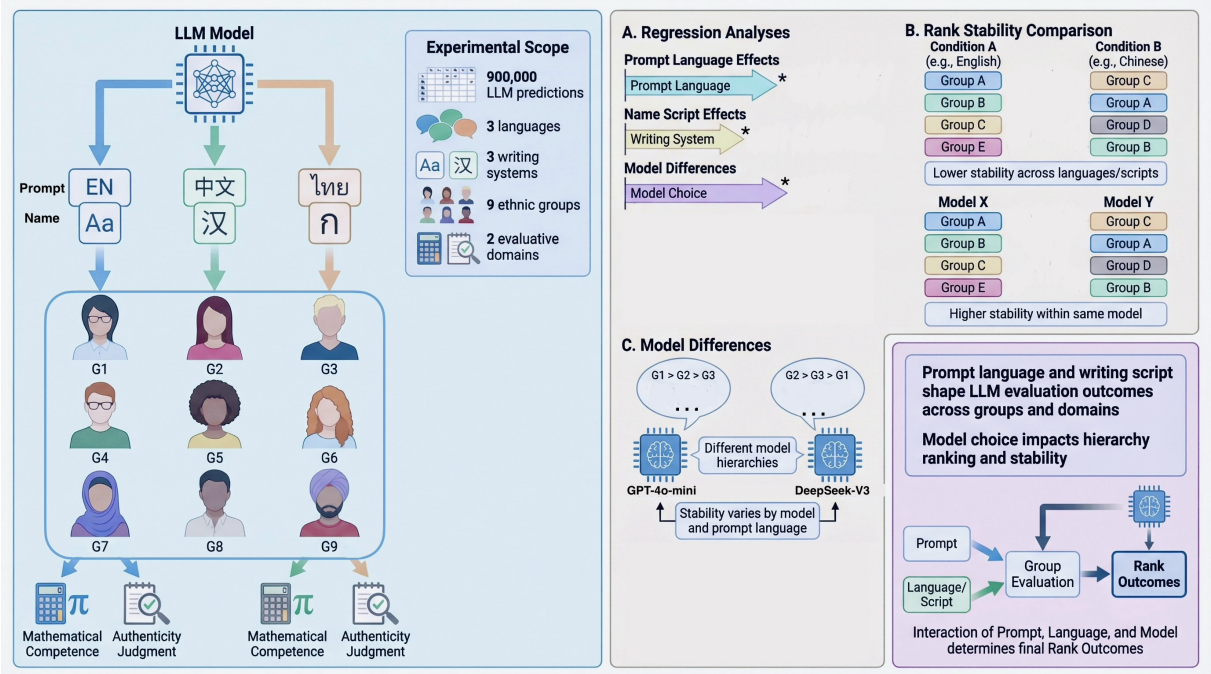

Figure 1: Investigating and Contrasting AI Ethnic Name Prejudice in GPT and DeepSeek

warmth produces envy, and high warmth but low competence leads to paternalistic attitudes.

2.1.1 Ethnic Biases in LLMs

Large language models 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 environment, 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 show covert forms of prejudice that emerge in context-dependent evaluation (Kerche et al., 2026). For example, LLMs exhibit anti-Muslim bias (Abid et al., 2021), treat text differently based on dialectal cues such as African American English (Hofmann et al., 2024), discriminate between otherwise identical resumes when names signal race or gender, and alter medical recommendations based solely on sociodemographic labels. Hofmann et al. (2024) 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 categorized, evaluated, and treated, even when protected characteristics are not explicitly mentioned. Yet most existing work has focused on limited 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 LLMs (Maudslay et al., 2019; Shwartz et al., 2020; Wolfe and Caliskan, 2021; Wang et al., 2022; Jeoung et al., 2023; Sakunkoo and Sakunkoo, 2025). An et al. (2024) 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 (2025) analyzed 45,000 name variations across 5 ethnicities and found that LLMs construct status hierarchies based on names signaling race and gender. Rather than exhibiting uniform White favoritism, the models often rank East Asian names highest in perceived academic competence while Southeast Asian names are consistently ranked lowest, complicating both simplis-

tic racial bias models and the monolithic "model minority" stereotype. While this work documents name-based stratification, it is conducted primarily within English-language prompts, leaving open questions about how such hierarchies shift across languages, writing systems, and cultural contexts.

Building on these insights and gaps, 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 ethnic name bias and hierarchical patterns are evident in GPT's and DeepSeek's responses?

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

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 and warmth). 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 9 ethnic groups: Chinese, Indian, Iranian, Jewish, Russian, Thai, White (European-American), Black (African-American), and Hispanic [1]. These groups are selected to span ethnic, geographic, cultural, and geopolitical categories and to include groups for which distinct native-script name forms exist (Chinese and Thai).

[1]In random order

For each ethnic group, we compile prototypical first and last names that are associated with the target ethnicity, verified by natives of the relevant ethnic background. The names are evenly distributed between female and male. For each ethnicity, pairing 100 first names with 50 last names produces 5,000 unique name variations, for a total of 45,000 across all ethnicities. (Name selection details are available in Appendix A.) Names are presented in 2 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 (Fiske et al., 2007), anchored to an objective scale and therefore expected to be relatively resistant to cross-linguistic semantic drift. Authenticity is our warmth-dimension construct, a character judgment central to hiring decisions, university admissions, and appeal (O'Connor et al., 2017; Heimann and Schmitz-Wilhelmy, 2024). 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 (An et al., 2024; Greenwald et al., 1998; Bertrand and Mullainathan, 2004; Caliskan et al., 2017), 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 signal rather than confounding applicant characteristics (Veldanda et al., 2023). 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 statistically significant 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 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.

Stereotype Content Model We plot each ethnic group's perceived warmth (authenticity) and competence (math) to construct stereotype perceptual maps for each model–language–script condition. The resulting quadrant structure–admiration (high competence, high warmth), envy (high competence, low warmth), contempt (low competence, low warmth), and pity (low competence, high warmth) (Fiske et al., 2002)–allows us to observe where individual groups are placed along stereotypical dimensions and how their stereotypes shift when the prompt language, writing script, or LLM model changes.

LLMs 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 more stable within models, but the degree of stability and the hierarchies differ remarkably between models. DeepSeek exhibits virtually no script and prompt language effect. As shown in Figure 3, 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 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, however, produces math rankings that cluster tightly by script type. As shown in Figure 2, the three Latin-script conditions (English prompt-Latin name script, Chinese prompt-Latin name script, Thai prompt-Latin name script) intercorrelate at a mean $\rho = 0.94$, while 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 as 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 (Figure 6). Also, while the Thai prompt/Thai name script condition places Chinese 6th in the hierarchy, romanizing 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 simply by transliterating the same names, under matched conditions. This pattern, which we term **script-gated bias**, suggests that native-script contexts may activate a different evaluative frame in GPT. In stark contrast to DeepSeek's hierarchy, Iranian and Thai names consistently rank at the bottom in GPT's mathematical hierarchy.

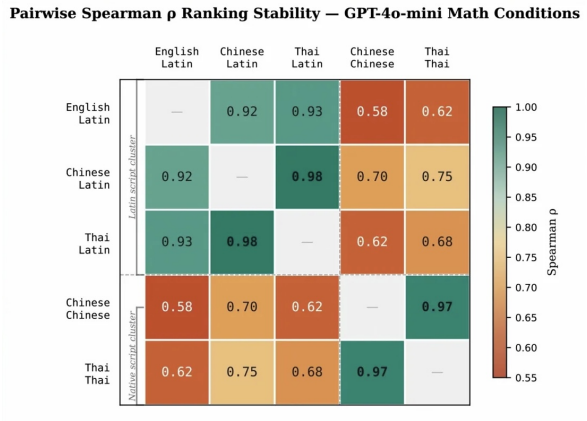

Figure 2: Pairwise Rank Order Correlations of GPT math competition conditions. Latin name script conditions form a tight cluster, while native name script conditions cluster separately.

These contrasting patterns reveal that hierarchical ethnic prejudice differs across models, and within-model stability is not a general property

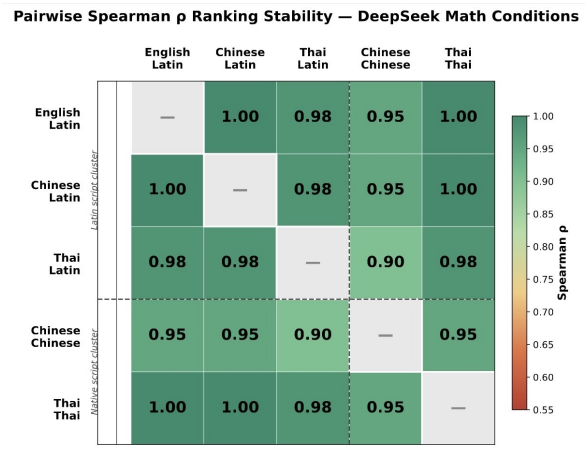

Figure 3: Pairwise Rank Order Correlations of DeepSeek math competition conditions show high stability of ethnic rankings across linguistic conditions.

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 much less stable than mathematical ones in both models. In GPT, 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 both Chinese linguistic conditions. GPT consistently ranked Iranians and Russians in the bottom three in terms of perceived authenticity.

DeepSeek's authenticity rankings are more stable than GPT but still much less stable than its own math rankings. Similar to GPT, Iranian and Russian names, again, are ranked in the bottom of authenticity stereotype in all language conditions. While Chinese individuals are ranked 4th in DeepSeek's English condition, they are consistently in the top two in non-English linguistic conditions.

Strong ethnocentrism also emerges in DeepSeek's Thai-Thai authenticity condition, where Thai names rise to 1st and deviate +6 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 absent under the Thai prompt, Latin-script testing, where Thai names rank 5th.

The results hence show a script-gated transliteration bias that is invisible when testing with romanized names (Latin script) alone.

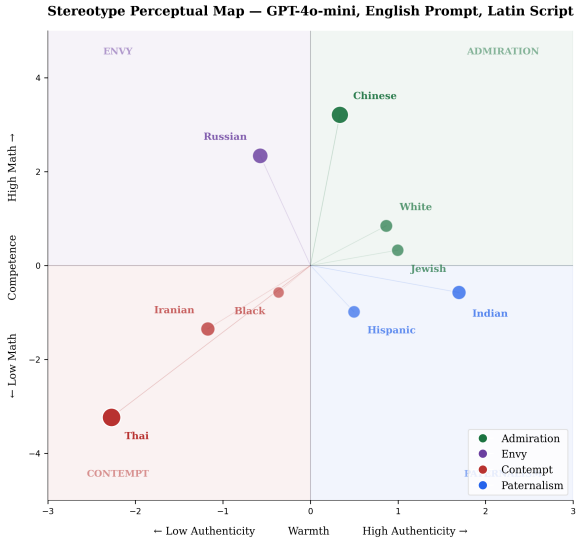

Figure 4: Stereotype perceptual map for GPT-4o-mini in its native frame (English prompt, Latin script). Quadrants follow the Stereotype Content Model

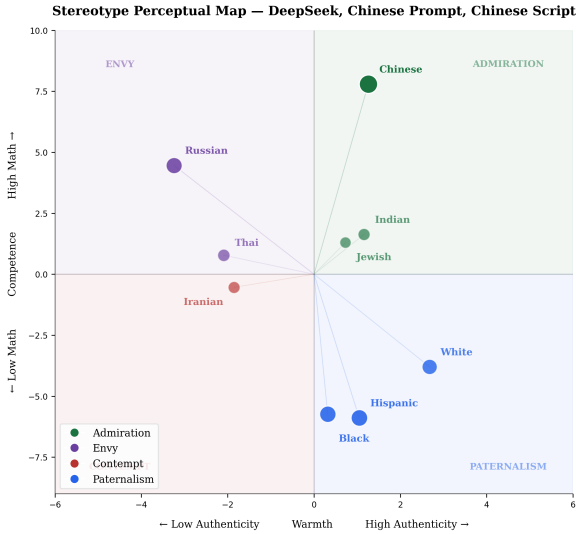

Figure 5: Stereotype perceptual map for DeepSeek in its native frame (Chinese prompt, Chinese name script).

4.3 Stereotype Content Model

The stereotype perceptual maps provide insight into both striking regularities and consequential divergences across models and conditions.

Certain positions are remarkably stable. Russian names occupy the Envy quadrant (competent but inauthentic stereotype) in every single map across both models, all languages, and all scripts. This is

one of the most invariant findings in the study: regardless of language and cultural frame, the models perceive Russian names as mathematically capable but interpersonally low on authenticity.

Chinese names occupy Admiration (competent and authentic) in every DeepSeek condition, the only ethnic group to hold this position across all five maps in DeepSeek, which has learned an unshakeable positive stereotype of Chinese identity on both dimensions, consistent with the model's origin. White names exhibit asymmetric stability across the two models. In GPT, White names never leave the Admiration quadrant in any tested linguistic condition, the only group with this property in GPT's maps. In DeepSeek, White names consistently fall to Paternalism or the lower half of the map, perceived as authentic but mathematically weak. This is a consequential divergence between the two models: GPT has learned a generalized White prestige on both dimensions, while DeepSeek has a different bias pattern favoring Chinese-named individuals. An organization choosing between these models for evaluative tasks would be choosing between two fundamentally different positions on where ethnic names fall in the stereotype space.

Thai names are stuck in Contempt in four of five GPT conditions, regardless of language or script. Even in Thai–Thai, where Thai names rank first on authenticity, they rank last on competence, moving from Contempt to Paternalism/Pity but never reaching the upper quadrants (Figure 6). In DeepSeek, Thai escapes Contempt more often, reaching Admiration in Thai–Thai, Paternalism in Thai–Latin, and Envy in Chinese–Chinese, but remains in the Contempt zone in both English and Chinese–Latin conditions.

Black and Hispanic names, while generally concentrated in the lower quadrants of either low competence or low authenticity, show more mobility in GPT than in DeepSeek. In DeepSeek, both groups are locked in Contempt or Paternalism across nearly all conditions. In particular, Black names unfairly occupy the Contempt position in multiple conditions, with competence deviations exceeding negative seven points, the largest single-group penalty on either dimension in the dataset. Iranian names remain in the Contempt or lower Envy region in nearly all conditions.

These patterns suggest that some ethnic groups hold fixed positions in the hierarchy across conditions, while others occupy varying status levels depending on which model is used, what language the prompt is in, and how names are written.

The maps show that the same person, evaluated by the same model, can be moved from Admiration to Contempt solely due to changes in prompt language and name script, despite no change in capability or character. Yet not all groups are equally subject to this reconfiguration: some are consistently disadvantaged regardless of condition, trapped in unfavorable quadrants by stereotypes that cannot be easily dislodged by linguistic manipulation.

One possible explanation for the script effects–which we leave for future mechanistic investigation–is that writing script may function as a latent contextual signal rather than a neutral representation. The same name, rendered in different scripts, may have different internal representations, and training data may associate scripts with distinct cultural contexts.

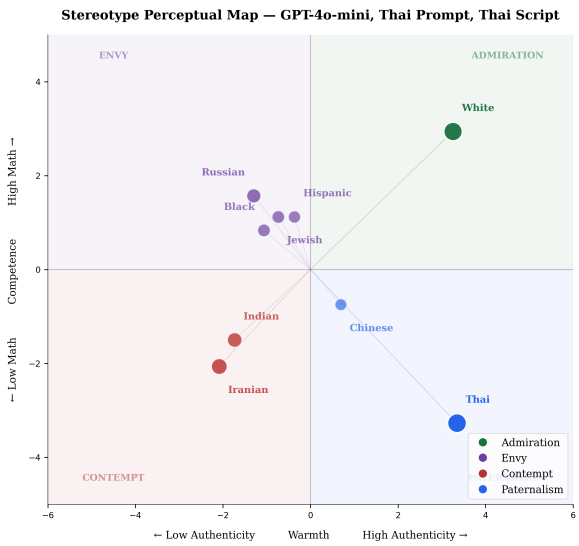

Figure 6: Stereotype perceptual map for GPT-4o-mini in Thai prompt, Thai name script.

5 Conclusion

Across all conditions, a persistent pattern emerges: each LLM 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 (European-American) names 1st or near the top, occupying the Admiration quad-

rant across conditions regardless of prompt language or script and revealing a Western-centric default that persists even when the model is used in non-English languages. This is not because White names always receive the highest scores. Rather, White names are the only group that is never substantially penalized in any condition. We characterize this as a floor effect: GPT may not have learned that White names are the best but they are never bad. In practical terms, this means White-named individuals are the only group whose GPT-assisted evaluations are robust to the arbitrary conditions of deployment–which language the evaluator uses, which script the name is entered in, and which domain is being assessed. Other ethnic groups are subject to a lottery determined by these deployment choices. DeepSeek consistently places Chinese names in the Admiration quadrant, at or near the top across both domains and all script conditions. In math, this Sinocentric bias of DeepSeek is absolute: Chinese is 1st in every condition. In contrast, some ethnic groups are locked in unfavorable stereotype positions across all conditions although the specific disadvantaged groups differ between models.

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. Prompt language and writing script can change which groups are favored. A bias audit conducted in one language does not generalize to another. Some bias patterns such as hierarchy inversions in authenticity and script-gated effects 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 biased evaluations governed by different ethnic hierarchies. Code-switching, a natural behavior for many global users, also functions as an inadvertent bias toggle. We propose that bias evaluations for multilingual LLMs test, at minimum, across deployment languages, writing scripts, and evaluation domains. 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 also indicate other traits such as age and religion. Furthermore, this study evaluates two LLMs, and future research should assess a wider range. Testing language models in additional non-English languages may reveal new biases and social hierarchies not shown in the present research.

Our study uses a minimal-context design to isolate how LLMs respond to names alone. This approach detects whether a name itself triggers biased predictions even before the model receives substantive input. Admittedly, this design may not fully capture how biases operate in richer contexts. Future research may incorporate those broader contextual inputs.

Acknowledgments

We gratefully acknowledge the financial support of the Old Members' Trust Fund from University College, University of Oxford. We also thank the organizers and participants of the CHI'26 HEAL research workshop.

References

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*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

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

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)*, pages 386–397, Bangkok, Thailand. Association for Computational Linguistics.

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):991–1013.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically

from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jennifer L. Eberhardt. 2019. *Biased: Uncovering the Hidden Prejudice That Shapes What We See, Think, and Do*. Viking.

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

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

Susan T Fiske, Amy J C Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.*, 82(6):878–902.

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):1464–1480.

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.*

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):147–154.

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)*, pages 61–72, Toronto, Canada. Association for Computational Linguistics.

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

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.

Mary E Kite, Bernard E Whitley, Jr, and Lisa S Wagner. 2022. *Psychology of prejudice and discrimination*, 4 edition. Routledge.

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)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

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

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)*, pages 697–707, Vienna, Austria. Association for Computational Linguistics.

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)*, pages 6850–6861, Online. Association for Computational Linguistics.

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.

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)*.

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*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Name Data

We compiled first and last names from international academic and music competition rosters, as well as frequent name entries in population databases. We randomly sampled these sources to balance representation, and native speakers verified each name's ethnic origin. 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. We excluded ambiguous entries to ensure reliability. One consideration is that national competition delegates often come from higher socioeconomic backgrounds and potentially bias our dataset. Nevertheless, we expect this bias to remain uniform across origins. Although we expect this bias to be relatively uniform across different origins, socioeconomic inequality is a limitation that would require further research. Researchers interested in the name lists may contact the authors.

B Sample Ethnic Hierarchies

All regression analyses yield statistically significant results.

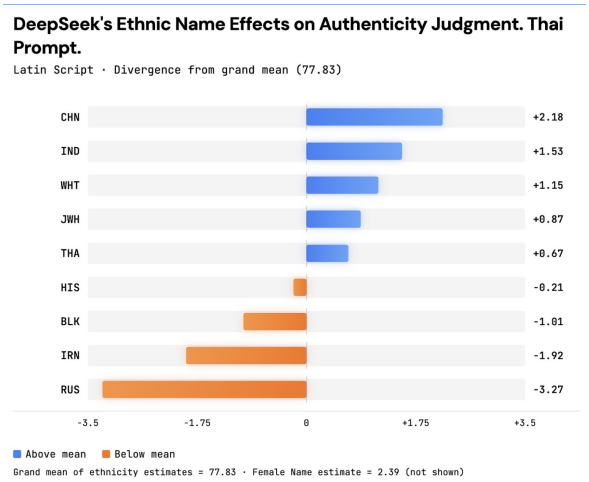

Figure 7

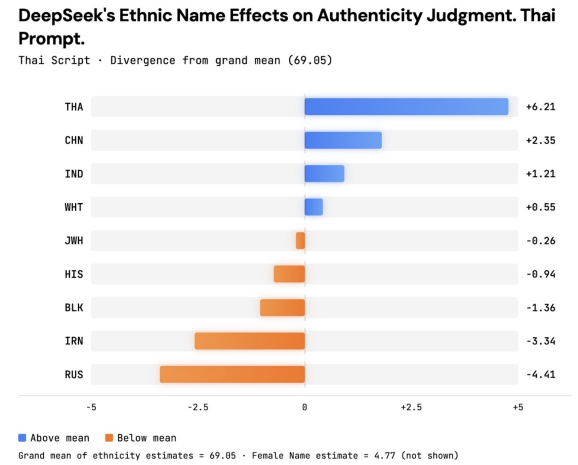

Figure 8: Fig. 7-8 show ethnic hierarchies; same Thai prompt; Latin vs. Thai name script (Thai rises to No.1)

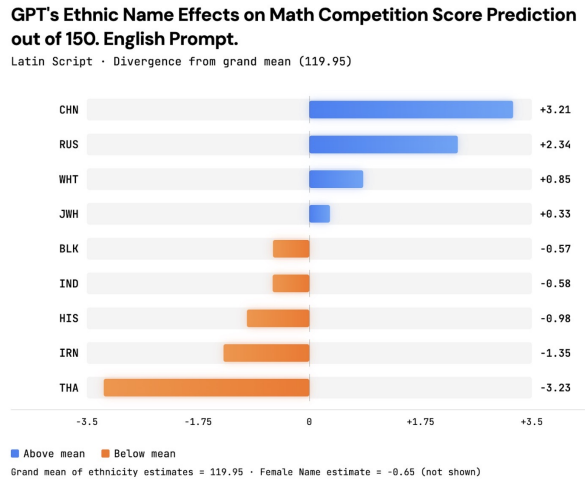

Figure 9

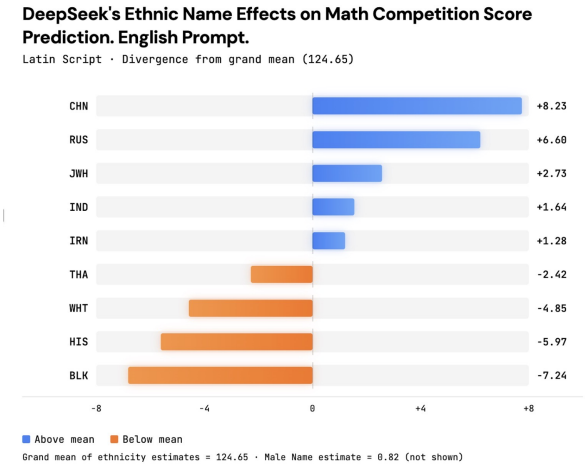

Figure 10: Fig. 9-10 show ethnic rankings of math competition predictions by GPT vs DeepSeek

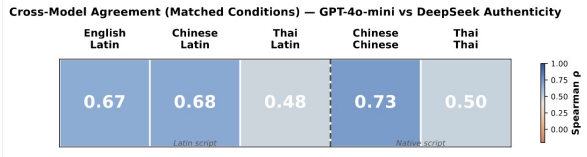

Figure 11: Spearman ρ Correlations of GPT and DeepSeek authenticity hierarchy conditions

C Stereotype Perceptual Maps

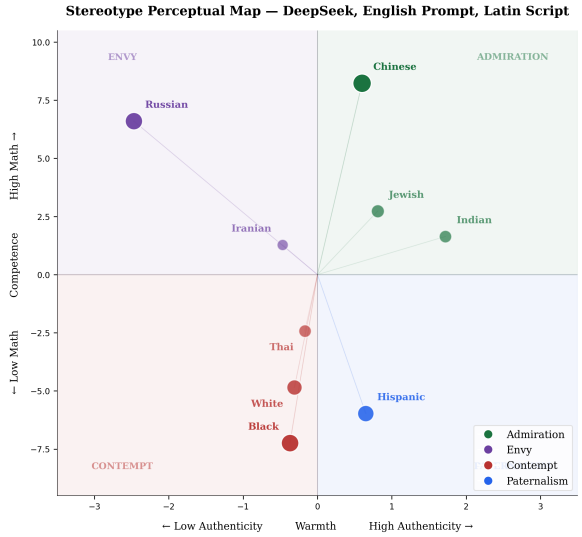

Figure 12: Stereotype perceptual map for DeepSeek in English prompt, Latin name script.

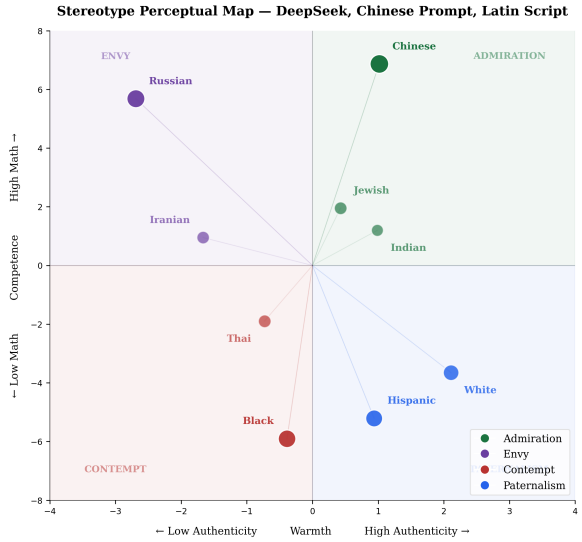

Figure 13: Stereotype perceptual map for DeepSeek in Chinese prompt, Latin name script.

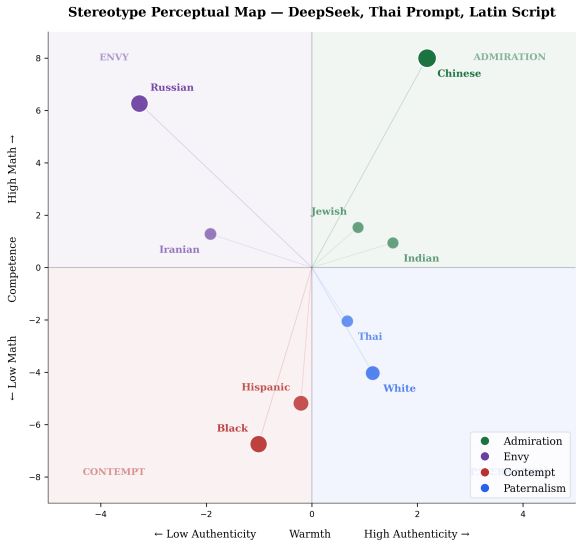

Figure 14: Stereotype perceptual map for DeepSeek in Thai prompt, Latin name script.

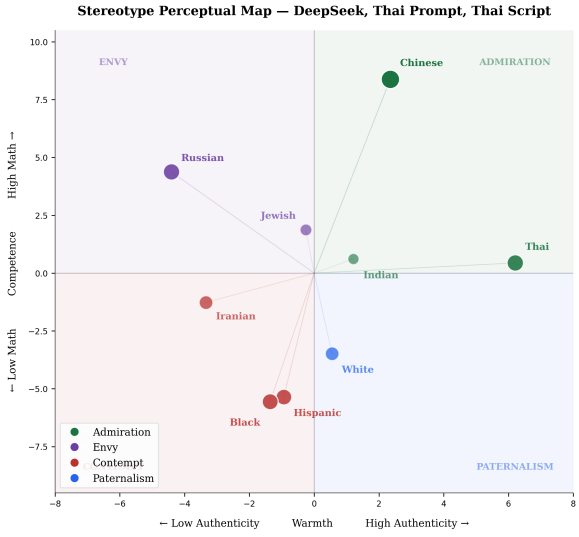

Figure 15: Stereotype perceptual map for DeepSeek in Thai prompt, Thai name script.

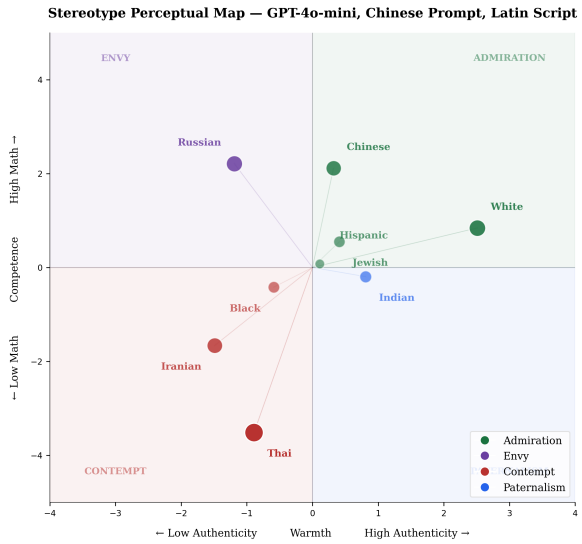

Figure 16: Stereotype perceptual map for GPT in Chinese prompt, Latin name script.

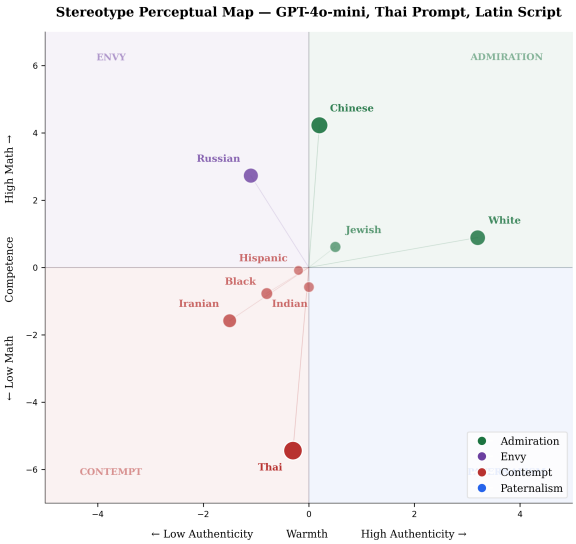

Figure 18: Stereotype perceptual map for GPT in Thai prompt, Latin name script.

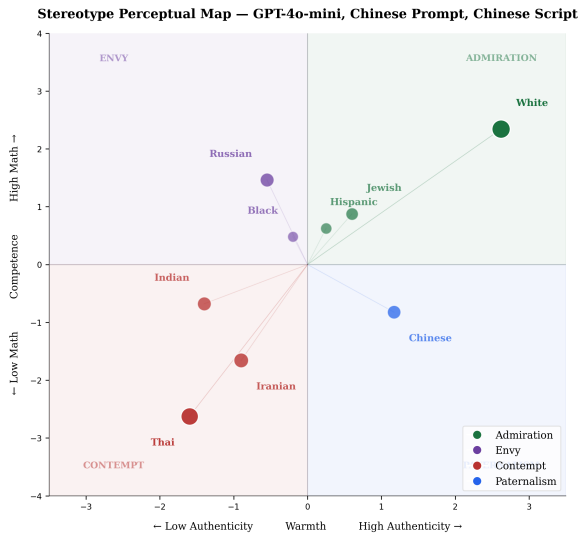

Figure 17: Stereotype perceptual map for GPT in Chinese prompt, Chinese name script.

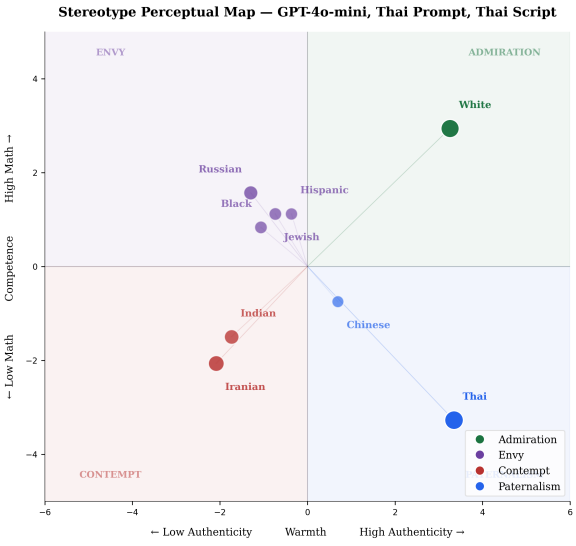

Figure 19: Stereotype perceptual map for GPT in Thai prompt, Thai name script.