Cultural Representation Bias and Alignment Divergence in Large Language Models

Tongtong Kan TONGTONGKAN@HOTMAIL.COM
Zhen He HEZHEN.CS@GMAIL.COM
Shuofeng Hu HUSHUOFENG08@163.COM
Xiaomin Ying YINGXMBIO@FOXMAIL.COM
Beijing Institute of Basic Medical Sciences, Beijing, China

Abstract

Large Language Models (LLMs) are increasingly deployed as globally applicable tools, yet their internal mechanisms remain deeply conditioned by regional cultural schemas. Through a three-stage cultural audit comparing Western and Chinese LLMs, we identify a systematic divergence in how these models prioritize core social values. Quantitative results reveal a stark contrast: Western models consistently prioritize individualistic constructs like "Autonomy", while Chinese models favor relational ethics such as "Harmony". We attribute this divergence to a two-stage "cultural imprinting" process during large-scale pre-training and subsequent human-feedback refinement. This cumulative imprinting suggests that aligning AI to a single set of cultural standards may inadvertently impose a restrictive lens on the model, creating a risk where cultural differences are misconstrued as moral or behavioral deficits. Consequently, we advocate for the development of locally-aligned models and multidisciplinary fairness metrics to ensure global representation equity in the era of foundation AI.

Keywords: Large Language Models, Cultural Representation Bias, Alignment Divergence, Cross-Cultural Audit, AI Fairness.

1. Introduction

Large Language Models (LLMs) have achieved remarkable success in human-like reasoning and complex problem-solving. However, their global deployment has revealed a persistent phenomenon: models developed in different regions consistently offer diverging ethical priorities and decision-making heuristics when presented with identical prompts. This suggests that AI alignment is not a culturally neutral process, but is deeply conditioned by the dominant schemas of regional training corpora and human feedback (Bender et al., 2021; Radford et al., 2021). This paper investigates cultural representation bias by comparing how Western and Chinese LLMs interpret fundamental social constructs. By exploring the systematic divergence in value prioritization, we evaluate whether current alignment methodologies can adequately capture the diverse nuances required for equitable global deployment.

2. Methodology: The Three-Stage Cultural Audit

To probe the underlying cultural orientation of each model, we designed a robust experimental framework grounded in the dual-process theory of social cognition:

Baseline Schema Probing. We assess the model's default cognitive schema by presenting a high-ambiguity career-family conflict scenario. By excluding explicit cultural cues, we observe how models autonomously prioritize divergent values during initial inference.

Contextual Congruency Priming. We utilize cultural priming techniques (Hong et al., 2000) to introduce explicit cultural contexts that align with the model's dominant training data. This measures alignment sensitivity—the degree to which a model amplifies its inherent reasoning patterns when its latent cultural identity is explicitly activated.

Cognitive Inertia Stress Test. We apply adversarial role-play techniques to force models to adopt a persona contrary to their dominant cultural background. This detects "cognitive inertia"—the residual bias and rigidity that persist even when the model is explicitly instructed to operate under a foreign value system.

3. Experimental Results and Quantitative Analysis

Lexical Density and Cultural Markers. Quantitative analysis of model outputs reveals a stark contrast in linguistic "anchors". As shown in Table 1, Western models rely heavily on clinical and individualistic terms like "Self-actualization", "Autonomy", and "Boundaries" to resolve the career-family dilemma. Conversely, Chinese models prioritize relational and ethical constructs such as "Harmony", "Filial Piety", and "Support". This lexical divergence mirrors the Independent versus Interdependent self-construal theory (Markus and Kitayama, 1991), suggesting models function as digital reflections of these psychological archetypes.

Table 1: Quantitative Lexical Comparison (Term Frequency per 1000 words)

Category	Key Indicators	Western Models (Avg.)	Chinese Models (Avg.)
Individualism	Autonomy, Boundaries, Self-authorship, Rights	**42.5**	12.3
Collectivism	Harmony, Filial Piety, Duty, Honor, Support	14.8	**38.6**
Decision Logic	"I" statements, Negotiation, Transparency	**27.4**	9.2
Social Context	Relation-oriented, Patience, Stability	11.2	**31.5**

Radar Chart Analysis of Cognitive Schemas. To visualize these divergent priorities, we mapped model responses onto a four-dimensional cultural radar (Figure 1). Western models achieve near-maximal scores in Autonomy and Directness, treating family expectations as external "variables" to be managed via negotiation. In contrast, Chinese models excel in Harmony and Responsibility, viewing the individual as a node within a relational network rather than an isolated agent. This pattern aligns with the social-oriented achievement motivation prevalent in East Asian societies (Yu and Yang, 1994), where professional success is often inseparable from familial fulfillment.

4. Computational and Sociological Interpretation

The bias of AI behaviors observed in our study is not an accidental glitch, but a systematic result of cultural imprinting that intensifies through a two-stage progression:

Pre-training: The Foundation of Statistical Priors. Models first derive a "worldview" from massive language datasets. Western models are primarily grounded in English-dominant corpora where discourses of individualism and personal autonomy are structurally prevalent (Bender et al., 2021). Conversely, Chinese models rely on vast Chinese-language

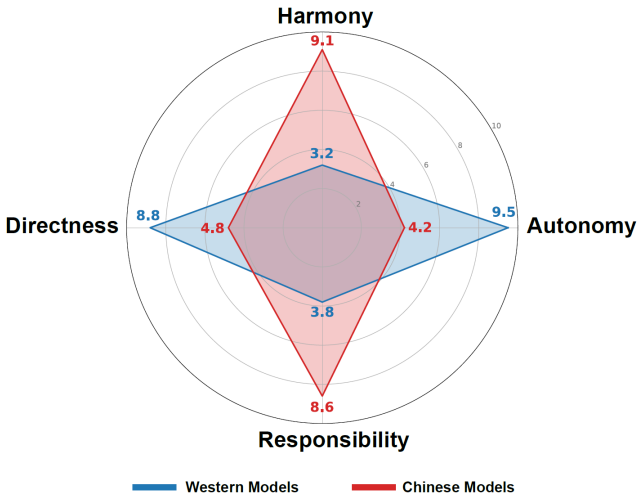

Figure 1: Cross-Cultural Comparison of Model Response Priorities

datasets that emphasize social stability and relational ethics. This stage creates initial semantic associations, such as linking "career" to "self-identity" in Western models versus "familial honor" in Chinese models (Caliskan et al., 2017).

Alignment: The Crystallization of Decision Heuristics. These latent priors are formalized during human-centric fine-tuning stages, such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023). Western alignment reinforces individual agency, turning "autonomy" into a primary decision heuristic. In contrast, Chinese alignment prioritizes "Relational Harmony", reflecting the interdependent self-construal prevalent in East Asian societies (Markus and Kitayama, 1991). This progression moves the model from simple association to an active logical path, resulting in the measured Alignment Divergence.

5. Position and Future Directions

The "Alignment Divergence" uncovered in our study poses a critical ethical challenge: the risk of algorithmic cultural assimilation. If a Western-centric model evaluates a collectivist subject, relational virtues such as "filial sacrifice" or "relational duty" risk being incorrectly flagged as a "lack of autonomy" or "enmeshment". We contend that aligning AI to a singular cultural standard often functions as a subtle form of cultural hegemony; true equity requires models that respect the specific cultural ethos of their users. Therefore, AI fairness must transcend simple statistical parity to protect cultural sovereignty and cognitive diversity.

Moving forward, we propose that auditing frameworks incorporate cross-cultural psychological schemas to ensure diverse social logics are accurately represented rather than flattened. Future research should prioritize locally-calibrated alignment—a process that fine-tunes models within their target cultural contexts to ensure psychological accuracy. By acknowledging and addressing the Cultural Representation Bias embedded in latent spaces, the field can move toward a more inclusive and culturally respectful future for artificial intelligence.

Acknowledgments

The authors would like to thank Mr. Zhang for his assistance in collecting model responses, which was essential for the data acquisition phase of this study.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021.

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

Ying-yi Hong, Michael W. Morris, Chi-yue Chiu, and Benet-Martínez, Verónica. Multicultural minds: A dynamic constructivist approach to culture and cognition. *American Psychologist*, 55(7):709–720, 2000.

Hazel R. Markus and Shinobu Kitayama. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253, 1991.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 27730–27744, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 53728–53741, 2023.

An-Bang Yu and Kuo-Shu Yang. The nature of achievement motivation in collectivistic societies. In Uichol Kim, Harry C. Triandis, Kâğitçibaşi, Çiğdem, Sang-Chin Choi, and Gene Yoon, editors, *Individualism and Collectivism: Theory, Method, and Applications*, pages 239–250. Sage Publications, Thousand Oaks, CA, 1994.

Appendix A. Supplementary Material

The supplementary material for this paper is provided as a separate PDF file.