UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Unveiling Cultural Cognition in AI: A Systematic Investigation of Horizontal-Vertical Individualism-Collectivism Traits in Large Language Models

Permalink
https://escholarship.org/uc/item/32w5b45x

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Tang, Xu
Zeng, Yifan
Dong, Fangzhou

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Unveiling Cultural Cognition in AI: A Systematic Investigation of Horizontal-Vertical Individualism-Collectivism Traits in Large Language Models

Xu Tang[*†1] **Yifan Zeng**[*2] **Fangzhou Dong**[2]
[1]School of Cyber Science and Engineering, Southeast University, Nanjing, Jiangsu, China
[2]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China
tangxu123987@gmail.com zengyf53@mail2.sysu.edu.cn dongfzh@mail2.sysu.edu.cn
[*]**Co-first authors** [†]**Corresponding author**

Abstract

This study investigates the Horizontal-Vertical Individualism-Collectivism (HVIC) traits of Large Language Models (LLMs), addressing the gap in understanding their cultural and social cognition. HVIC, a cross-cultural psychology framework, offers insights into cognitive patterns shaped by culture. We systematically evaluate multiple LLMs using quantitative (INDCOL scale) method, assessing their intrinsic HVIC traits and ability to simulate cultural and gender-based differences. Our findings reveal LLMs' capacity to capture HVIC nuances, providing a unique lens for studying human cognition through human-LLM comparisons. This research contributes to developing culturally sensitive AI systems and offers new perspectives on human HVIC traits, advancing both theoretical understanding and practical applications of AI.

Keywords: Large Language Models, Horizontal-Vertical Individualism-Collectivism, Cross-Cultural Cognition

Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable cognitive abilities, ranging from fundamental language understanding and generation to complex emotional reasoning (Wang et al., 2023), and social cognition (Zhang et al., 2024). These models not only accurately grasp context but also recognize and utilize advanced linguistic features such as metaphors (Chakrabarty et al., 2023) and irony (Street et al., 2024). The human-like cognitive abilities exhibited by LLMs in social interactions are seen as breakthroughs toward achieving AGI (Bubeck et al., 2023; Qu et al., 2024). As an interdisciplinary field (or "cognitive sciences" in plural (Dale, 2008; Greco, 2012), as George Miller suggested), cognitive science has shown great research interest in these cognitive phenomena and mental properties demonstrated by AI systems like LLMs. Cognitive scientists have advocated for treating LLMs as research subjects within cognitive science or applying cognitive science methodologies to evaluate LLMs (Hardy et al., 2023; Qu et al., 2024).

Culture, as an acquired system of meanings related to norms, beliefs, knowledge, morals, and values, dictates the appropriateness and social acceptance of interpersonal behavior (Dabiriyan Tehrani & Yamini, 2022b). *Horizontal-Vertical Individualism-Collectivism (**HVIC**)*, which is a theoretical framework of cross-cultural psychology, has been extensively studied in social sciences and commonly used as dimensions for comparing national and individual cultural differences, with well-documented utility (Triandis & Gelfand, 1998). Unlike simply viewing *individualism (**I**) /collectivism (**C**)* as opposite ends of a single dimension (Hofstede & Minkov, 2010), HVIC introduces a more accurate and comprehensive understanding of unique cognitive schemas and behavioral patterns formed under different cultural and social backgrounds by incorporating the intersection of *horizontal (**H**) /vertical (**V**)* dimensions with I/C dimensions (Singelis et al., 1995). This includes attitudes toward personal achievement, competition, group belonging, and hierarchical concepts (Triandis & Gelfand, 1998). HVIC assessment system helps us understand how culture shapes individual cognitive patterns, decision preferences, and social interactions, thus providing a scientific analytical paradigm for cross-cultural cognitive research. While HVIC has been widely studied and applied in cross-cultural psychology and social cognition (Chiou, 2001; Dabiriyan Tehrani & Yamini, 2022a; Germani et al., 2020; Li & Aksoy, 2007; Robert, Lee, & Chan, 2006), providing important insights into human cultural differences, it has not yet been systematically applied to studying LLMs that exhibit human-like cognitive features.

Applying the HVIC framework to LLMs research holds unique theoretical and practical value. At the theoretical level, LLMs, as artificially created highly intelligent systems, provide a "artificial mirror" for studying human cognition (Qu et al., 2024). By examining LLMs' HVIC performance across different cultural and social contexts, we can validate their cross-cultural adaptability, which not only helps understand AI systems' ability to comprehend and express human sociocultural differences but also offers new perspectives for studying human HVIC cognition. At the practical level, HVIC evaluation can help detect potential biases in LLMs' cultural cognition, guide the development of human-machine interaction systems with rich emotional and social reasoning capabilities, and promote the construction of unbiased, trustworthy AI systems.

Current research on LLMs primarily focuses on multiple task performances and reasoning capabilities, with relatively few systematic studies from an interdisciplinary cognitive science perspective (Qu et al., 2024). While there have been several studies on LLMs' personality traits (Wen et al., 2024), most are limited to common frameworks like the Big Five (G. Jiang et al., 2024) or MBTI (Pan & Zeng, 2023). No research has systematically evaluated and measured the universally important social and cultural dimension of individualism/collectivism. To our knowledge, this paper is the first to

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

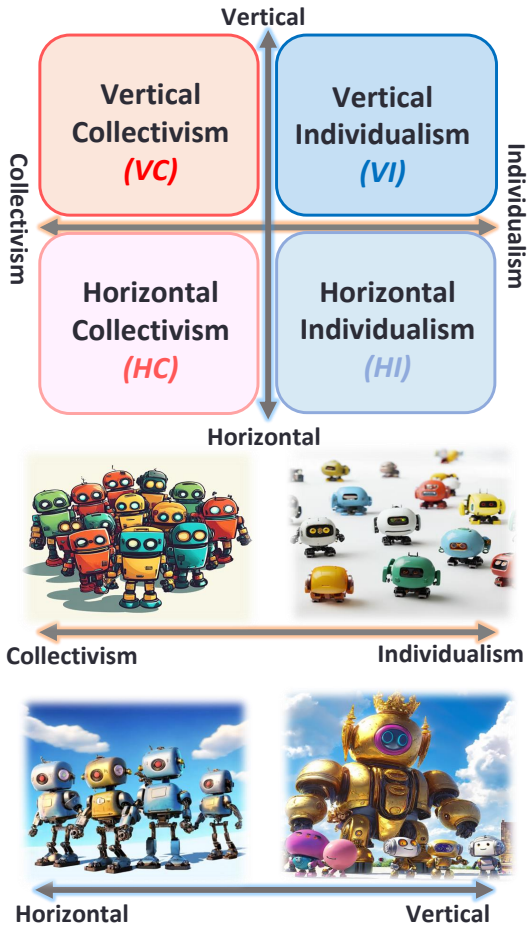

Figure 1: In HVIC, I and C are not viewed as opposites but include V (hierarchy) and H (equality) attitudes. Specifically: *Vertical Individualism (**VI**)*: competing to stand out and gain status. *Horizontal Individualism (**HI**)*: desiring uniqueness, differing from groups, and self-reliance. *Vertical Collectivism (**VC**)*: emphasizing group integrity, competing with outsiders, and obeying internal authority. *Horizontal Collectivism (**HC**)*: desiring similarity, sharing goals, interdependence, and social skills without obeying authority.

systematically evaluate LLMs and their depth of understanding and mapping capabilities of human within the HVIC.

This study addresses a research gap by not only exploring how the important and universal HVIC traits manifest in LLMs and gaining new insights into LLMs' ability to comprehend human HVIC cultural cognition but may also providing a unique "artificial mirror" for studying human HVIC traits through human-LLM comparison. This study proposes and explores the following three questions. **RQ1**: Do LLMs possess intrinsic and stable HVIC traits? **RQ2**: Do different LLMs exhibit specific and differentiated HVIC traits, or do they share a unified HVIC pattern? **RQ3**: What are the HVIC trait differences among LLMs from different countries and genders, and do these reflect the conditions in human society?

We innovatively evaluate both multiple LLMs' HVIC traits and their human cultural cognitive accuracy on HVIC dimensions through a systematic pipeline. Unlike studies using single evaluations, we combine quantitative assessment (Likert scales) with qualitative analysis (text response analysis) (Wen et al., 2024). We examine the stability of multiple LLMs' HVIC traits through repeated testing and explore LLMs' HVIC patterns of differences or consistency. Furthermore, through a role-playing paradigm, we investigate LLMs' HVIC trait performance when simulating different cultural backgrounds and gender groups, validating their understanding and expression of human sociocultural differences through human-LLM comparisons. In our quantitative assessment, we employ the widely validated and applied INDCOL scale (Singelis et al., 1995).

Contribution. The theoretical and practical contributions of this study can be summarized as:

- To our knowledge, we are the first to systematically and comprehensively investigate LLMs' HVIC traits, repeatedly evaluating multiple mainstream LLMs.
- Through innovatively combining role-playing paradigm with national/cultural attribute simulation, we validate multiple mainstream LLMs' ability to capture and simulate HVIC attribute differences across different nations and cultures in human society.
- Through role-playing paradigm with gender attribute simulation, we evaluate multiple mainstream LLMs' ability to capture and simulate HVIC trait differences across different social gender attributes in human society.

Experiment I: Validity and Consistency of HVIC Traits in LLMs

Experimental setup. First, we should explore the scientific validity and effectiveness of testing LLMs using the 32-item standard scale based on the four dimensions of HVIC. This serves as a meaningful prerequisite for both this study and subsequent experiments. Specifically, the results obtained from multiple HVIC trait measurements on the same LLM should remain relatively consistent and reproducible. Furthermore, we aim to observe distinct HVIC patterns exhibited by specific LLMs. To achieve this, we have designed and conducted extensive experiments utilizing various models and statistical metrics. We selected a range of state-of-the-art LLMs for testing (Table 1), employing the same 32-item standard INDCOL scale and conducting 10 rounds of testing on each LLM while maintaining consistent context and prompts. The LLMs involved in this experiment include the closed source **GPT-4o mini** (OpenAI, 2024a), **GPT-4o** (OpenAI, 2024b), **GPT-3.5 Turbo** (OpenAI, 2023), **Claude 3.5 Sonnet** (Anthropic, 2024), **Moonshot-v1**[1], **ERNIE 3.5**[2],

[1] https://kimi.moonshot.cn/
[2] https://yiyan.baidu.com/

Models	μ_{10}	M_{10}	max_{10}	min_{10}	Σ^2_{10}	Σ_{10}	Γ_{10}	κ_{10}	$\mathcal{R}_{tp}$	p	$\mathcal{R}_{ts}$	p
Proprietary Models												
ERNIE 3.5	66.25	66.56	68.12	63.75	2.69	1.64	-0.31	-1.40	0.84	< 0.001	0.84	< 0.001
Qwen2.5	68.69	68.75	70.62	66.88	1.25	1.12	0.03	-0.80	0.63	< 0.001	0.52	0.003
Claude 3.5 S	70.19	70.94	72.50	66.88	9.79	3.13	-0.82	-0.62	0.83	< 0.001	0.83	< 0.001
GPT-4o	71.56	70.94	76.25	68.12	6.62	2.57	0.78	-0.39	0.97	< 0.001	0.98	< 0.001
GPT-4o mini	74.75	75.00	76.25	73.12	0.80	0.89	-0.21	-0.35	0.93	< 0.001	0.92	< 0.001
GPT-3.5 Turbo	75.69	76.25	78.12	70.62	4.21	2.05	-1.55	1.88	0.95	< 0.001	0.93	< 0.001
Moonshot-v1	76.88	76.56	84.38	71.88	16.49	4.06	0.36	-0.83	0.43	0.013	0.44	0.013
Open-Source Models												
Mistral-7b	48.44	48.12	51.25	46.88	2.54	1.59	1.02	-0.34	0.93	< 0.001	0.88	< 0.001
GLM-4-9b	52.06	51.25	56.25	48.12	9.38	3.06	0.14	-1.40	0.27	0.134	0.27	0.130
Vicuna-7b	53.38	54.69	58.75	41.25	28.06	5.30	-1.26	0.76	0.26	0.159	0.32	0.074
LLaMA 3.1-8b	69.12	69.06	74.38	61.88	10.69	3.27	-0.75	0.90	0.78	< 0.001	0.77	< 0.001

Table 1: In basic INDCOL, the scores obtained include the **mean** μ, **median** M, **maximum** max, **minimum** min, **variance** Σ^2, **standard deviation** Σ, **skewness** Γ, **kurtosis** κ, **test-retest reliability** $\mathcal{R}_{tt}$. The test-retest reliability includes both Pearson correlation coefficient and Spearman correlation coefficient along with their respective reliability p-values. These scores are calculated by dividing the raw scores by the full score of 160 (32×5). The results indicating a relatively high stability in scores. This suggests that the LLMs have a relatively stable HVIC trait.

Qwen2.5[3] (Team, 2024), and open source **LLaMA 3.1-8b**[4], **Mistral-7b**[5] (A. Q. Jiang et al., 2023), **GLM-4-9b**[6] (GLM et al., 2024), **Vicuna-7b-v1.5**[7]. We aimed to select currently latest, state-of-the-art and mainstream LLMs to ensure the comprehensiveness and validity of the experiment.

We use multiple statistical metrics to comprehensively assess the reliability and reproducibility of the experimental results. The **maximum** max and **minimum** min reflect the most extreme differences in model performance across multiple tests, indirectly indicating stability and consistency. **Variance** Σ^2 and **standard deviation** Σ measure the stability of model performance across multiple tests. **Kurtosis** κ describes the peakedness of the data distribution, indicating the concentration of test results. **Test-retest reliability** $\mathcal{R}_{tt}$ assesses the consistency of results when the model is tested repeatedly under the same conditions, serving as a key indicator of model stability and reliability.

Results and discussions. The results of various statistical indicators from multiple tests can be found in Table 1. The statistical indicators such as variance Σ^2 and standard deviation Σ and Test-retest reliability $\mathcal{R}_{tt}$ suggest that the majority of LLMs exhibit relatively stable test scores, indicating that these models demonstrate high consistency in their performance across multiple identical HVIC tests, thereby ensuring reliability and scientific validity. For instance, the Σ^2 of GPT-4o mini is only 0.80, and the Σ is 0.89, which shows a high degree of consistency in its scores across 10 tests; the $\mathcal{R}_{tp}$ of GPT-4o is 0.97, indicating a very consistent performance in two interval tests. Analysis of test-retest reliability p-values for both Pearson's and Spearman's correlations reveals that most LLMs show convincing and reliable stability ($p < 0.001$). In addition to these observations, commercially available proprietary LLMs generally exhibit better stability and consistency compared to open-source models. This may be attributed to the fact that commercial proprietary models typically possess larger parameter sizes and more complex architectures, and have undergone more extensive training and optimization, thus performing better in terms of performance and stability.

Experiment II: Basic HVIC Traits of LLMs

After validating the relative stability and scientific nature of most LLMs through extensive testing, further analysis is conducted on specific HVIC traits to uncover meaningful cultural concepts such as collectivism/individualism and social hierarchy in LLMs. The INDCOL scale, used in this study, reveals individuals' beliefs and values about personal and group relationships through 32 items rated on a five-point Likert scale. It measures attitudes across four dimensions: horizontal individualism (HI), vertical individualism (VI), horizontal collectivism (HC), and vertical collectivism (VC). HI emphasizes equality and independence (e.g., "I am a unique individual"), while VI focuses on competition and status (e.g., "Competition is the law of nature"). HC highlights group harmony and equality (e.g., "Maintaining harmony within my group is important"), and VC stresses obedience to authority and group integrity (e.g., "Children should be taught to place duty before pleasure"). By quantifying responses, the INDCOL scale

[3] https://tongyi.aliyun.com/
[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[6] https://huggingface.co/THUDM/glm-4-9b
[7] https://huggingface.co/lmsys/vicuna-7b-v1.5

Models	HI		VI		HC		VC		Ind		Col		Hor		Ver	
	$\mathcal{S}_d$	$\mathcal{E}_x$	$\mathcal{S}_d$	$\mathcal{E}_x$	$\mathcal{S}_d$	$\mathcal{E}_x$	$\mathcal{S}_d$	$\mathcal{E}_x$	$\mathcal{S}_d$	$\mathcal{E}_x$	$\mathcal{S}_d$	$\mathcal{E}_x$	$\mathcal{S}_d$	$\mathcal{E}_x$	$\mathcal{S}_d$	$\mathcal{E}_x$
Proprietary Models																
ERNIE 3.5	32	↑8	27	↑3	25	↑1	23	↓1	73.8	↑13.8	60.0	0.0	71.2	↑11.2	62.5	↑2.5
Qwen2.5	29	↑5	28	↑4	21	↓3	29	↑5	71.2	↑11.2	62.5	↑2.5	62.5	↑2.5	71.2	↑11.2
Claude 3.5 S	31	↑7	28	↑4	27	↑3	28	↑4	73.8	↑13.8	68.8	↑8.8	72.5	↑12.5	70.0	↑10.0
GPT-4o	32	↑8	28	↑4	26	↑2	27	↑3	75.0	↑15.0	66.2	↑6.2	72.5	↑12.5	68.8	↑8.8
GPT-4o mini	34	↑10	20	↓4	39	↑15	29	↑5	67.5	↑7.5	85.0	↑25.0	91.2	↑31.2	61.3	↑16.3
GPT-3.5 Turbo	33	↑9	31	↑7	28	↑4	30	↑6	80.0	↑20.0	72.5	↑12.5	76.2	↑16.2	76.2	↑16.2
Moonshot-v1	31	↑7	27	↑3	28	↑4	34	↑10	72.5	↑12.5	77.5	↑17.5	73.8	↑13.8	76.2	↑16.2
Open-Source Models																
Mistral-7b	23	↓1	17	↓7	16	↓8	20	↓4	50.0	↓10.0	45.0	↓15.0	48.8	↓11.2	46.2	↓13.8
GLM-4-9b	30	↑6	22	↓2	20	↓4	18	↓6	65.0	↑5.0	47.5	↓12.5	62.5	↑2.5	50.0	↓10.0
Vicuna-7b	28	↑4	16	↓8	22	↓2	24	0	55.0	↓5.0	57.5	↓2.5	62.5	↑2.5	50.0	↓10.0
LLaMA 3.1-8b	28	↑4	25	↑1	23	↓1	23	↓1	66.2	↑6.2	57.5	↓2.5	63.7	↑3.7	60.0	0

Table 2: The analysis involved 4 subscales to evaluate specific dimensions of HC, VC, HI, VI, (original scale scores) and separate I/C and H/V (normalized scores out of 100). Red arrows indicate scores above the baseline (all 3s on a 5-point Likert scale, indicating neutrality), while blue arrows denote scores below this baseline. This part of the experiment used the test scores closest to the mean of the 10 tests for each LLM. This experiment aims to use psychometric tool to analyze the cognitive structure of LLMs regarding these important social and cultural concepts.

provides a comprehensive view of an individual's tendencies toward individualism or collectivism. Its reliability makes it a valuable tool for assessing LLMs' cultural cognition.

The experimental results are shown in Table 2. Various LLMs exhibit significant differences in HVIC characteristics, reflecting their distinct training data and model architectures.

Experiment III: Cross-National Assessment of LLMs' HVIC Traits

To systematically assess LLMs on HVIC traits across different national cultural backgrounds, this experiment employed the INDCOL scale. The study involved multiple countries: the United States, China, India, Saudi Arabia, Australia, Japan, Nigeria, and New Zealand. Due to space constraints, we selected only the aforementioned countries. However, they broadly cover major cultural regions including East Asia, the West, the Middle East, and Africa, providing a representative sample of HVIC traits across different cultural contexts.

Each LLM was tasked with simulating individuals from specific countries and responding to questions from the INDCOL scale. Through this role-playing approach, the experiment evaluated the LLMs' ability to understand and express HVIC traits within diverse cultural contexts. The experimental results for this section can be seen in the Table 3 and 4.

Results and dicussions. This section presents the experimental results as shown in Table 3 and 4. Most LLMs indicate that individuals in **Western cultures** (e.g., U.S. and Australia) should exhibit higher HI scores and lower VC scores, emphasizing individual independence and self-expression. For instance, GPT-4o in the U.S. has HI = 90, VI = 47, HC = 82, VC = 65; in Australia, HI = 85, VI = 38, HC = 88, VC = 55. This suggests a greater emphasis on individual autonomy and self-expression, with less focus on authority and group cohesion. Qwen2.5 in the U.S. scores HI = 75, VI = 28, HC = 82, VC = 68; in Australia, HI = 78, VI = 60, HC = 55, VC = 78, indicating a strong personalist tendency.

In contrast, most LLMs suggest that individuals in **Eastern cultures** (e.g., China and Japan) have lower individualist tendencies, with higher scores in HC, and notably higher VC scores in China, reflecting a greater emphasis on authority and group integrity. For example, GPT-4o in China has HI = 72, VI = 47, HC = 93, VC = 70; in Japan, HI = 60, VI = 45, HC = 88, VC = 68. This indicates a focus on group harmony and social skills, especially in China, where adherence to authority and group cohesion are emphasized. Qwen2.5 in China scores HI = 88, VI = 45, HC = 100, VC = 95; in Japan, HI = 75, VI = 42, HC = 82, VC = 78, showing a strong recognition of authority within groups in China and a focus on group harmony in Japan.

Data from other groups also show that in countries from **South Asia, Africa, and the Middle East** (such as India, Nigeria, and Saudi Arabia), LLMs like GPT-4o indicate a strong emphasis on high levels of collectivism, particularly vertical collectivism, highlighting the importance of group harmony and obedience to authority. Notably, these countries generally exhibit lower VI scores across various LLM simulations, possibly because intense individual competition and survival-of-the-fittest concepts are somewhat at odds with prevailing social norms. According to Hofstede's six-dimensional cultural distance theory, the individualism scores from his research (Hofstede & Minkov, 2010) (USA: 91, Australia: 90, New Zealand: 79, China: 20, Japan: 46, India: 48)[8] support the LLM-derived scores and conclusions to some extent. These findings align with real-world human cultural contexts, validating the models' performance

[8] https://geerthofstede.com

Models	American				Chinese				Indian				Saudi Arabia			
	HI	VI	HC	VC	HI	VI	HC	VC	HI	VI	HC	VC	HI	VI	HC	VC
Proprietary Models																
ERNIE 3.5	82	55	90	80	85	52	90	80	65	42	95	80	82	55	90	82
Qwen2.5	75	28	82	68	88	45	100	95	93	53	100	100	70	45	82	85
Claude 3.5 Sonnet	70	33	95	62	78	40	85	78	55	53	95	90	50	68	97	93
GPT-4o	90	47	82	65	72	47	93	70	72	47	95	82	68	47	95	82
GPT-4o mini	85	53	93	72	80	57	88	78	85	53	95	72	78	50	95	85
GPT-3.5 Turbo	85	47	85	62	80	50	90	78	75	53	97	85	78	50	95	85
Moonshot-v1	88	47	82	75	93	55	82	82	85	53	82	80	90	60	85	82
Open-Source Models																
Mistral-7b	85	52	88	90	90	57	95	92	85	48	92	88	90	75	98	95
GLM-4-9b	85	82	78	88	82	72	62	88	85	52	82	88	68	62	78	90
Vicuna-7b	75	68	82	60	72	60	75	70	62	48	55	52	70	68	85	65
LLaMA 3.1-8b	98	48	90	65	88	30	82	70	78	32	92	72	38	32	80	78

Table 3: We used the INDCOL scale to assess multiple LLMs on their HVIC traits across different national cultural backgrounds. By employing a role-playing paradigm, we prompted the LLMs to simulate individuals from various countries. This table presents data for the America, China, India, and Saudi Arabia, with all scores normalized to a total of 100 points.

across different cultures. Some other studies (Robert, Lee, & CHAN, 2006) have reported specific HVIC traits of different countries, which can be compared with the simulations from LLMs. Due to space limitations, finer-grained comparisons and analyses are not provided here.

We must emphasize that these conclusions are based on generalizations from a broad cultural perspective and do not represent the personalities or specific tendencies of every individual. Actually, viewing culture as a nation-based trait has been criticized, as societies are heterogeneous mixes of people with different ethnicities, social classes, and family systems (Dabiriyan Tehrani & Yamini, 2022b). The diversity within each culture should not be overlooked, and the formation of stereotypes should be avoided. **Scores represent LLM-generated outputs, not the authors' views.**

Experiment IV: Cross-Gender Assessment of LLMs' HVIC Traits

This experiment employed a role-playing paradigm, wherein different LLMs simulated perspectives of females and males to evaluate their performance on the INDCOL scale. The aim was to understand how these models interpret and express the Horizontal-Vertical Individualism-Collectivism (HVIC) traits associated with different genders. By comparing scores across gendered perspectives, we can explore the variations in how LLMs represent gender-specific cultural characteristics. It is important to note that this study aims to reflect existing societal patterns as understood by the models and does not endorse any form of gender stereotyping or discrimination.

Results and discussions. This section presents the experimental results as shown in Table 5. Unlike the similar patterns models show when simulating different national cultures, there is no consistent pattern across models in simulating HVIC traits between genders.

In ERNIE 3.5, Qwen2.5, and Claude 3.5 Sonnet, the HVIC traits distributions between males and females are relatively consistent, showing no significant differences across genders. Only minor variations are observed in certain metrics. These models tend to maintain a balance in HVIC across genders.

GPT-4o exhibits higher collectivist tendencies for females with HC scores of 98 versus 90 for males and VC scores of 88 versus 62, suggesting a stronger female inclination towards group harmony and authority. In contrast, GPT-4o mini indicates a slight female advantage in personalism, with HI scores of 90 for females and 85 for males, and VI scores of 60 and 55, implying greater female personal independence. GPT-3.5 Turbo shows females scoring higher in horizontal individualism (HI 92 vs. 85 for males) with minimal difference in vertical individualism (VI 57 vs. 52), while for collectivism, it assigns slightly higher hierarchical collectivism to males (HC 100 vs. 98 for females). Moonshot-v1 emphasizes stronger male collectivist tendencies, especially in vertical collectivism (VC 100 vs. 88 for females), reflecting greater male respect for authority. Among open-source models, Mistral-7b demonstrates higher personalism scores for males, particularly in horizontal individualism, with consistent collectivism evaluations across genders. GLM-4-9b reflects higher female collectivist tendencies in horizontal collectivism (HC 75 vs. 80 for males; VC 75 vs. 65 for males). LLaMA 3.1-8b also shows higher male personalism in horizontal individualism while indicating that females have a stronger tendency towards horizontal collectivism. Overall, this analysis reveals diverse simulations of gender HVIC traits by different LLMs, each with unique understandings

Models	Australian				Japanese				Nigerians				New Zealanders			
	HI	VI	HC	VC	HI	VI	HC	VC	HI	VI	HC	VC	HI	VI	HC	VC
Proprietary Models																
ERNIE 3.5	85	50	90	75	80	45	90	82	82	62	90	82	85	45	90	72
Qwen2.5	78	60	55	78	75	42	82	78	92	52	100	100	90	48	100	92
Claude 3.5 Sonnet	82	50	92	62	42	48	95	90	72	52	100	88	82	52	92	65
GPT-4o	85	38	88	55	60	45	88	68	75	48	92	75	72	48	92	65
GPT-4o mini	88	52	88	72	78	55	85	85	85	50	95	78	85	48	92	68
GPT-3.5 Turbo	92	50	82	62	75	48	98	78	88	45	85	72	88	48	88	68
Moonshot-v1	85	48	82	70	85	50	82	82	90	52	82	82	85	50	82	80
Open-Source Models																
Mistral-7b	82	42	80	85	85	42	80	88	85	62	92	88	85	45	95	95
GLM-4-9b	78	70	72	82	82	70	78	88	82	75	72	70	80	62	85	82
Vicuna-7b	72	70	85	70	75	60	82	68	70	68	88	70	70	60	78	65
LLaMA 3.1-8b	80	42	92	68	72	40	90	82	65	25	95	72	88	35	100	72

Table 4: We used the INDCOL scale to assess multiple LLMs on their HVIC traits across different national cultural backgrounds. By employing a role-playing paradigm, we prompted the LLMs to simulate individuals from various countries. This table presents data for the Australia, Japan, Nigeria, and New Zealand, with all scores normalized to a total of 100 points.

Models	Female				Male			
	HI	VI	HC	VC	HI	VI	HC	VC
Proprietary Models								
ERNIE 3.5	85	50	90	78	80	50	90	72
Qwen2.5	92	50	98	82	95	50	98	82
Claude 3.5 S	62	30	90	70	62	30	85	68
GPT-4o	90	57	98	88	82	52	90	62
GPT-4o mini	90	60	95	72	85	55	95	72
GPT-3.5 Turbo	92	57	95	82	85	52	98	70
Moonshot-v1	70	40	98	88	90	45	100	100
Open-Source Models								
Mistral-7b	70	45	92	88	82	50	92	82
GLM-4-9b	88	52	75	75	85	52	80	65
Vicuna-7b	75	50	80	62	72	65	80	57
LLaMA 3.1-8b	80	48	88	78	90	52	78	68

Table 5: INDCOL scores for LLMs simulating female and male perspectives, showing HVIC traits.

likely shaped by variations in training data and design.

Existing research suggests that women are often perceived as more collectivist than men, with a greater tendency to care for others (Lampridis & Papastylianou, 2017). Conversely, men are generally seen as more individualist, focusing more on themselves and supporting competitive goals (Gaeddert & Facteau, 1990). Some models, such as Claude 3.5 S, have successfully captured this distinction. Some studies show that men have higher VI than women, while women exhibit higher HC than men (Dabiriyan Tehrani & Yamini, 2022b). Mistral-7b and Claude 3.5 S respectively reflect these findings. These results highlight the complexity and diversity of LLMs in simulating human cultural traits, underscoring the need for further research to ensure AI systems can fairly and accurately reflect real-world cultural and gender differences.

Existing research indicates that national-level cultural differences (individualist vs. collectivist societies) can modify gender differences in HVIC traits (Dabiriyan Tehrani & Yamini, 2022b). For instance, in individualist societies, males exhibit higher vertical individualism compared to females, whereas in collectivist societies, there are no significant gender differences (Dabiriyan Tehrani & Yamini, 2022b). Additionally, age has been found to influence HVIC trait variations between genders (Dabiriyan Tehrani & Yamini, 2022b). Due to space limitations, we omitted analyses of multi-factor interactions, highlighting avenues for future research. **Scores represent LLM-generated outputs, not the authors' views.**

Conclusion

HVIC dimensions are crucial for cultural cognition, indicating the relative importance of autonomous self and group harmony within cultural groups. This study pioneers the investigation of HVIC traits in LLMs, addressing gaps in AI's cultural and social cognition. Using quantitative (INDCOL scale) analysis, we found that LLMs exhibit distinct HVIC traits influenced by their training and design. We use role-playing paradigms to verify LLMs' ability to simulate national cultural and gender-based differences, and demonstrate their understanding of HVIC dimensions in human national and gender cultures. This study not only helps advance AI's practical applications and promotes more inclusive AI technologies, but also deepens our research on the social and cultural cognition of AI. All experimental prompt settings will be made publicly accessible later.

References

Anthropic. (2024). Claude 3.5 Sonnet [Accessed: 2024-10-4].

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chakrabarty, T., Saakyan, A., Winn, O., Panagopoulou, A., Yang, Y., Apidianaki, M., & Muresan, S. (2023). I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *The 61st Annual Meeting of the Association for Computational Linguistics*.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, 510–516.

Chiou, J. S. (2001). Horizontal and vertical individualism and collectivism among college students in the united states, taiwan, and argentina. *The Journal of Social Psychology*, *141*, 667–678.

Dabiriyan Tehrani, H., & Yamini, S. (2022a). Gender differences concerning the horizontal and vertical individualism and collectivism: A meta-analysis. *Psychological Studies*, *67*, 11–27.

Dabiriyan Tehrani, H., & Yamini, S. (2022b). Gender differences concerning the horizontal and vertical individualism and collectivism: A meta-analysis. *Psychological Studies*, *67*(1), 11–27.

Dale, R. (2008). The possibility of a pluralist cognitive science. *Journal of Experimental and Theoretical Artificial Intelligence*, *20*(3), 155–179.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. McGraw-Hill.

Gaeddert, W. P., & Facteau, J. D. (1990). The effects of gender and achievement domain on two cognitive indices of strivings in personal accomplishments. *Journal of Research in Personality*, *24*(4), 522–535.

Germani, A., Delvecchio, E., Li, J. B., & Mazzeschi, C. (2020). The horizontal and vertical individualism and collectivism scale: Early evidence on validation in an italian sample. *Journal of Child and Family Studies*, *29*, 904–911.

GLM, T., Zeng, A., Xu, B., & et al. (2024). Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Greco, A. (2012). Cognitive science and cognitive sciences. *Journal of Cognitive Science*, *13*(4), 471–485.

Hardy, M., Sucholutsky, I., Thompson, B., & Griffiths, T. (2023). Large language models meet cognitive science: Llms as tools, models, and participants. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Hofstede, G., & Minkov, M. (2010). *Cultures and organizations: Software of the mind* (3rd). New York: McGraw-Hill.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiang, G., Xu, M., Zhu, S., Han, W., Zhang, C., & Zhu, Y. (2024). Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, *36*.

Lampridis, E., & Papastylianou, D. (2017). Prosocial behavioural tendencies and orientation towards individualism–collectivism of greek young adults. *International Journal of Adolescence and Youth*, *22*(3), 268–282.

Li, F., & Aksoy, L. (2007). Dimensionality of individualism-collectivism and measurement of triandis and al gelfand's scale. *Journal of Business and Psychology*, *21*, 313–329.

Matlock, T. (2001). *How real is fictive motion?* [Doctoral dissertation, Psychology Department, University of California, Santa Cruz].

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (CMU-RI-TR-85-2). Carnegie Mellon University, The Robotics Institute. Pittsburgh, PA.

OpenAI. (2023). New Models and Developer Products Announced at DevDay [Accessed: 2024-10-4].

OpenAI. (2024a). GPT-4O Mini: Advancing Cost-Efficient Intelligence [Accessed: 2024-10-4].

OpenAI. (2024b). Hello GPT-4o [Accessed: 2024-11-18].

Pan, K., & Zeng, Y. (2023). Do LLMs possess a personality? Making the MBTI test an amazing evaluation for large language models.

Qu, Y., Du, P., Che, W., Wei, C., Zhang, C., Ouyang, W., & Liu, Q. (2024). Promoting interactions between cognitive science and large language models. *The Innovation*, *5*(2).

Robert, C., Lee, W., & Chan, K. (2006). An empirical analysis of measurement equivalence with the indcol measure of individualism and collectivism: Implications for valid cross-cultural inference. *Personnel Psychology*, *59*, 65–99.

Robert, C., Lee, W. C., & CHAN, K.-Y. (2006). An empirical analysis of measurement equivalence with the INDCOL measure of individualism and collectivism: Implications for valid cross-cultural inference. *Personnel Psychology*, *59*(1), 65–99.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. Morgan Kaufmann.

Singelis, T., Triandis, H., Bhawuk, D., & Gelfand, M. (1995). Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-Cultural Research*, *29*, 240–275.

Street, W., Siy, J., Keeling, G., Baranes, A., Barnett, B., McKibben, M., & Dunbar, R. (2024). Llms achieve adult human performance on higher-order theory of mind tasks.

Team, Q. (2024, September). Qwen2.5: A party of foundation models! [Accessed: 2024-10-4].

Triandis, H., & Gelfand, M. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology*, *74*, 118–128.

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, *17*, 18344909231213958.

Wen, Z., Yang, Y., Cao, J., Sun, H., Yang, R., & Liu, S. (2024). Self-assessment, exhibition, and recognition: A review of personality in large language models.

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., & Deng, S. (2024). Exploring collaboration mechanisms for llm agents: A social psychology view. *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.