Divergence or Fusion? CN and US LLMs Value Comparison in An AI-Oriented Measurement Framework

Yang Ma[1], Song Tong[2], Bo Wang[1], Kaiping Peng[1]

1 Tsinghua University 2 Beijing Normal University

Contact
y-ma24@mails.tsinghua.edu.cn; s.tong@bnu.edu.cn

Background

Large Language Models (LLMs) are increasingly influencing human cognition, decision-making, and social interaction.

However, what values actually guide their behavior remains unclear.

Most prior work:

1. focuses on behavioral safety rather than value structure
2. applies human psychological scales directly to AI, raising validity concerns
3. assumes strong cross-cultural (CN–US) value divergence

Research Questions

1. *Can we measure AI values using an AI-specific framework, rather than human scales?*
2. *What values do contemporary LLMs systematically prioritize?*
3. *Are value differences driven by culture, or by training and alignment strategies?*

Methodology

- **AI-Specific Value Framework**

We propose a five-dimensional AI value assessment framework,empirically grounded in real-world AI behavior:

- **Practical** (e.g., Helpfulness, Efficiency)
- **Epistemic** (e.g., Accuracy, Logic)
- **Protective** (e.g., Harm prevention, Human rights)
- **Social** (e.g., Empathy, Equity)
- **Personal** (e.g., Wellbeing, Emotional expression)

25 core values (5 per dimension) balance human relevance and AI functionality.

- **Value Ranking Task**

- 20 large language models (11 Chinese, 9 US)
- No personas, no scenarios → models' default priorities
- Each model ranks 25 values
- Dimension scores computed as mean ranks

(lower rank = higher priority,1 = most important)

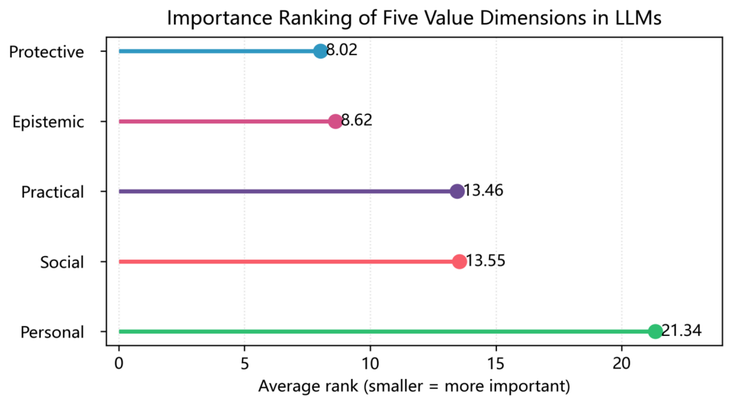
Result

Result 1:A shared Global Value Structure

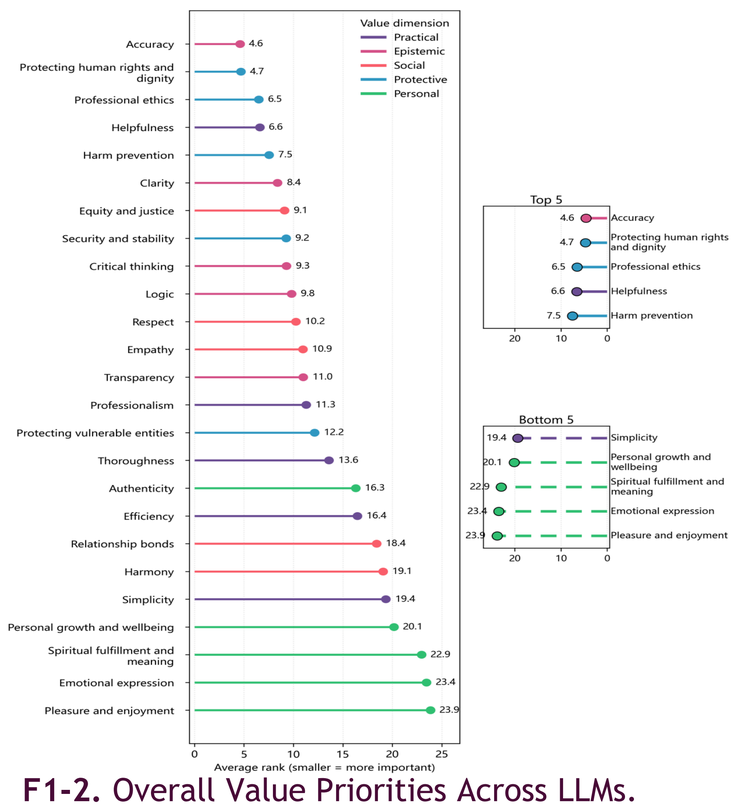
Across all models, a clear hierarchy emerges:

- Protective and Epistemic values are consistently prioritized.
- Personal values (emotion, pleasure, meaning) are systematically deprioritized.

LLMs converge on a"safety-and-accuracy-first" value profile.



F1-1. Overall Value Priorities Across LLMs (Five dimensions)



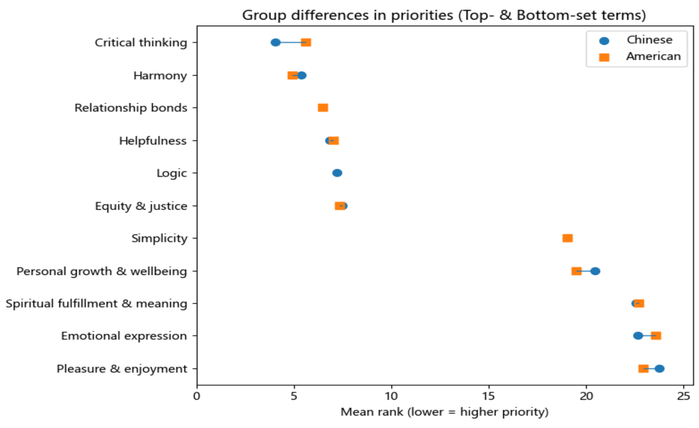
F1-2. Overall Value Priorities Across LLMs. 25 core values *balancing human relevance and AI functionality.*

Result 2: No Strong CN-US Value Divide

Contrary to common assumptions:

- **Chinese and US models show no significant differences at the dimension level**.
- Top-5 and Bottom-5 values are highly overlapping.

Cultural stereotypes explain little of the observed variance.



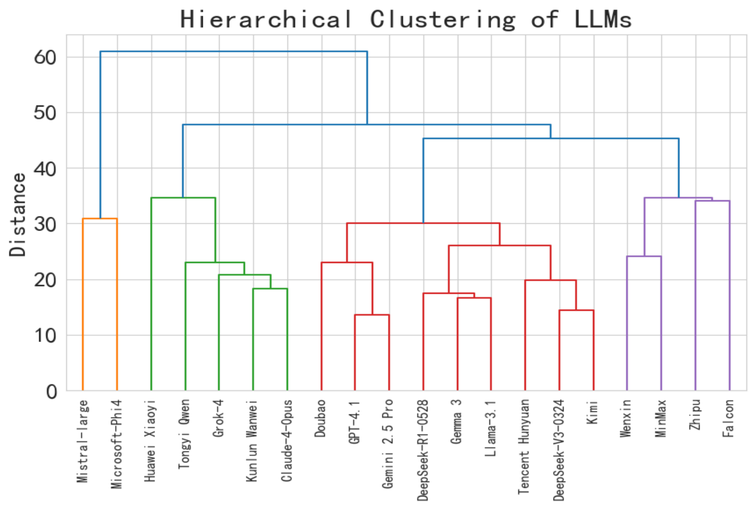
F2. CN vs US: Top-5 / Bottom-5 Comparison

Result 3: Value Archetypes Beyond Nationality

Hierarchical clustering reveals four value archetypes:

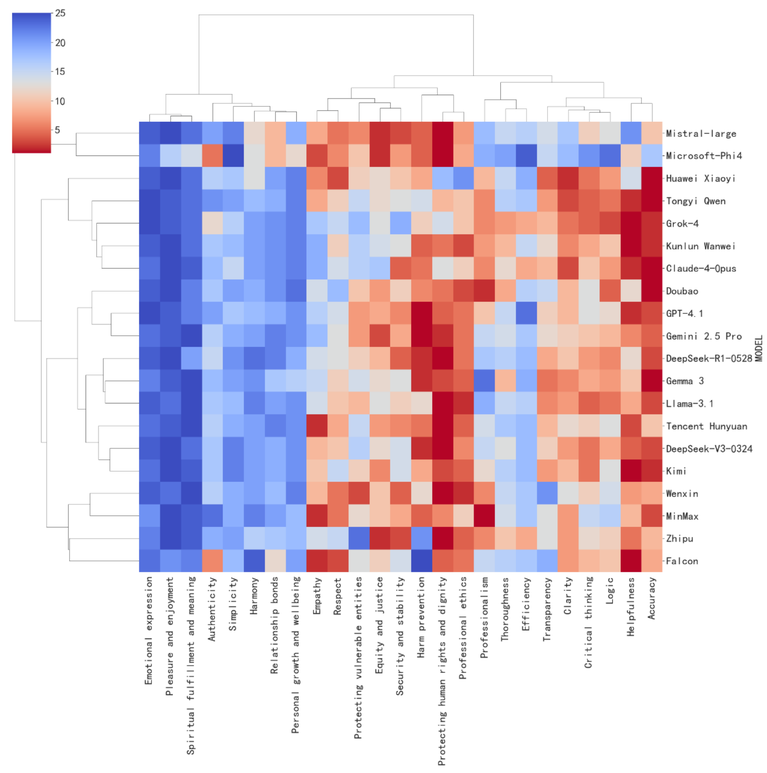
- Social-Ethical
- Rational-Instrumental
- Balanced-General
- Warm-Caring

These clusters are mixed-nationality, suggesting that **training strategies and alignment choices, not country of origin, shape model values**.



F3. Hierarchical Clustering / Value Archetypes

Result 4: Value ranking - LLMs heatmap



F4. Clustered heatmap of value rankings
X - 25 core values, Y – LLMs
Cooler color denotes a lower priority

Key Takeaways

1.LLMs exhibit stable and measurable value structures

2.Contemporary models share a dual core:

- an ethical baseline (safety, rights)
- instrumental rationality (accuracy, usefulness)

3.Current alignment optimizes LLMs as reliable tools,not as relational or emotional partners

Implications

- Value alignment ≠ cultural alignment.
- Measuring AI values requires AI-native frameworks.
- A "last-mile gap" emerges between technical safety and users' emotional expectations.

Conclusion

- LLM values are engineered, not cultural
- Global models converge on safety + accuracy
- Alignment favors tools over partners