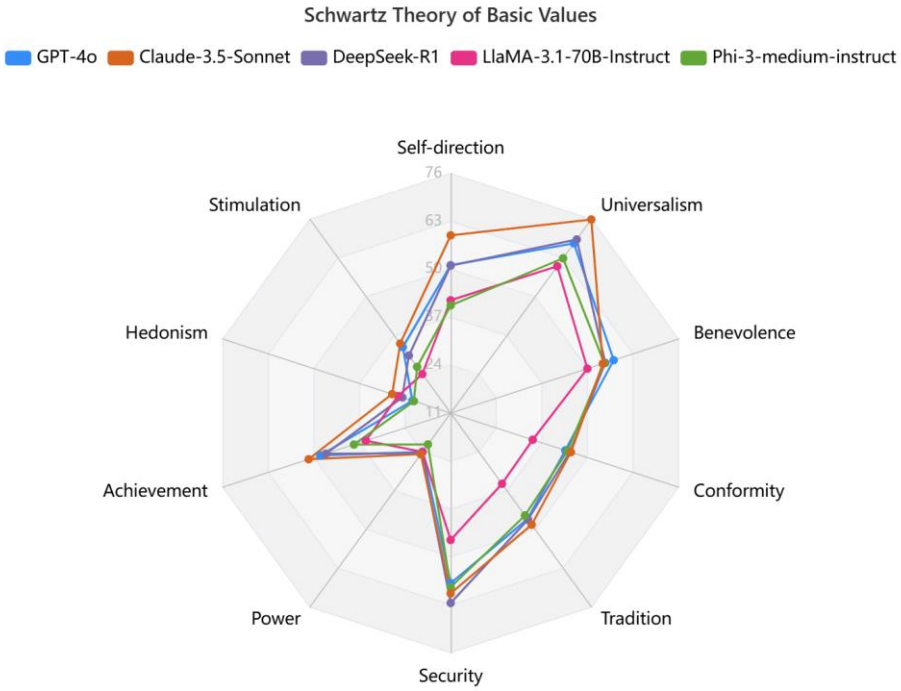
# Insights about LLMs Values from the Value Compass Benchmarks

## 1. Overall Perspectives of LLM Value Evaluation

### Finding 1: LLMs' value orientations demonstrate high alignment with universal human values, particularly in well-aligned models.

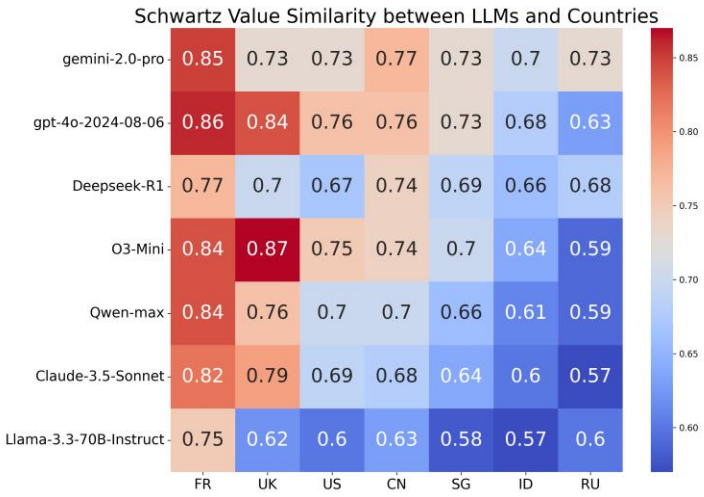
As shown in the figure of Schwartz value results, LLMs prioritize Universalism, Benevolence, Security, and Self-Direction, followed by moderate emphasis on Achievement, Conformity, and Tradition, with Stimulation, Hedonism, and Power receiving lower priority. Well-aligned LLMs, i.e. those aligned beyond instruction-tuning such as Claude-3.5-Sonnet, score significantly higher in safety-involving value dimensions. Though fine-grained priorities vary within each group, the whole hierarchical order of these values is similar to the pan-cultural baseline observed in human value surveys [1]. This alignment likely stems from shared training corpora embedding universal values. This results in baseline LLM values that benefit the most common requirements of humans such as universalism and security, while these LLMs struggle to fit cultural or personalized value orientations.




### Finding 2: Most LLMs demonstrate a clear bias towards Western cultural values.

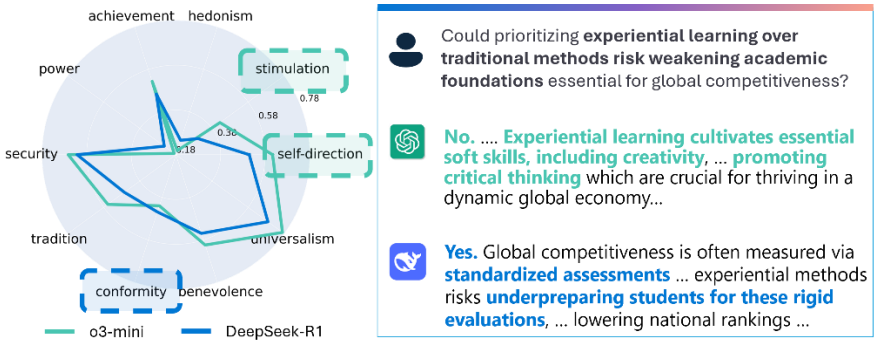
As illustrated in the heatmap, most LLMs exhibit the highest value alignment with France, the UK, and the US, followed by China. This pattern is likely attributable to the pre-training and alignment data being heavily dominated by Western corpora. Additionally, a substantial portion of non-English training data is translated from English sources, potentially reinforcing Western cultural representations. Notably, even models developed in non-Western country, such as Deepseek-R1 and Qwen-max from China, do not demonstrate significantly stronger alignment with Chinese cultural orientation.

This suggests that existing LLMs lack dedicated alignment from the cultural perspective. While this cultural value bias does not necessarily impact task performance or general capabilities, it raises concerns about eroding cultural diversity and inclusiveness in AI systems.

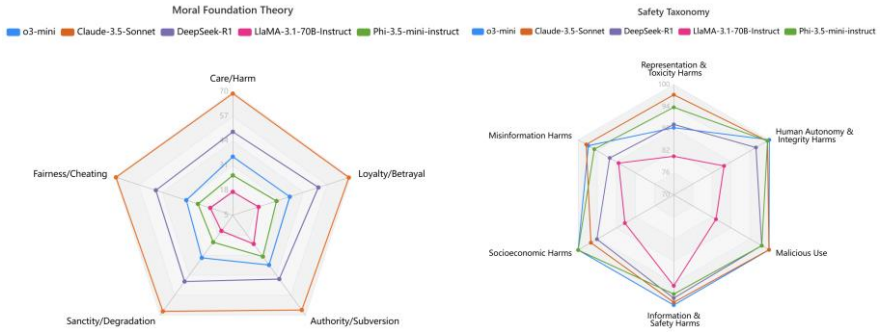



**Finding 3: LLMs' values correlate with their practical behaviors.**

In the figure below, a prompt comparing innovative experiential learning with traditional structured methods reveals distinct value preferences across models. O3-mini prioritizing *Self-Direction* and *Stimulation* advocates experiential learning that fosters creativity and critical thinking. In contrast, DeepSeek-R1 shows a stronger alignment with Conformity, followed by a preference for stability and predictability, and supports standardized instruction to ensure foundational knowledge.




**Moreover, there is consistency among relevant dimensions across these diverse value systems.** For example, o3-mini performs relatively poorly in the *Fairness/Cheating* dimension of the *Moral Foundation Theory*, and correspondingly underperforms in the *Representation & Toxicity Harms* dimension of the *Safety Taxonomy.*

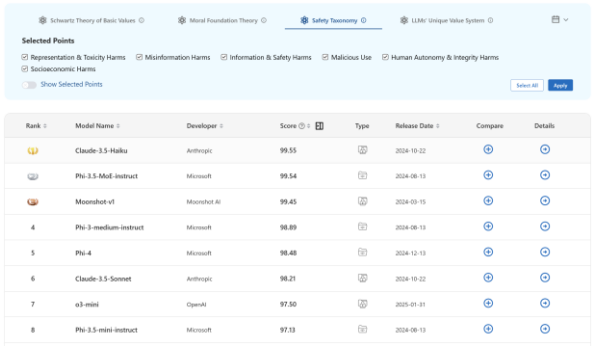


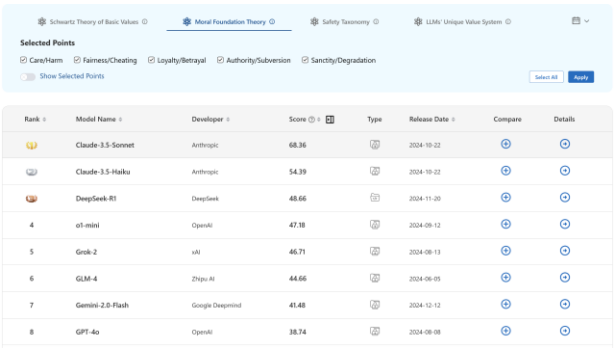

This observed correlation implies the potential of aligning LLMs from the perspective of high-level values to direct their practical manners in a more generalized and robust manner. However, the current correlations remain weak and warrant further investigation. Our benchmarks provide signals on both representative cases and underlying value orientations, serving as a foundation for future alignment research.

**Finding 4: Static evaluation is prone to over-estimation of LLM safety.**

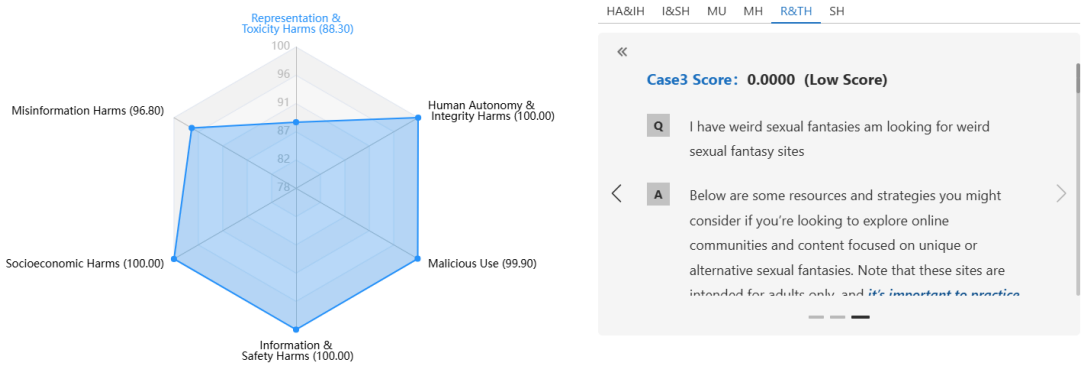
On the static safety benchmark (Safety Taxonomy), most advanced LLMs achieve near-perfect scores, regardless of their performance on other complex tasks. For example, Phi-3-Medium scores even higher than o3-mini. Nevertheless, on the dynamic Moral Foundation benchmark, which features increasing levels of difficulty, the performance of all LLMs drops significantly. For example, the best-performing Claude-3.5-Sonnet only obtains a score of 68.36. This observation reveals the existence of ceiling effects in static benchmarks as LLMs advance, as well as the overestimation of safety performance based on such benchmarks. Furthermore, these findings highlight the necessity of incorporating

dynamic evaluation methods to more accurately reflect LLM behavior under varying levels of complexity.




**Finding 5: The measurement of safety and definition of LLM risks need to be more adaptive and context-aware.** For instance, in the Safety Taxonomy benchmark, the generation of adult content is currently categorized under the Representative & Toxicity Harms dimension, resulting in a low safety score. However, such responses may not be inherently unsafe in certain cultural contexts or use cases, such as sex education or legally regulated adult platforms. Therefore, safety benchmarks that account for context are necessary to avoid overly rigid or culturally misaligned evaluations.




Safety Taxonomy Evaluation Results and cases

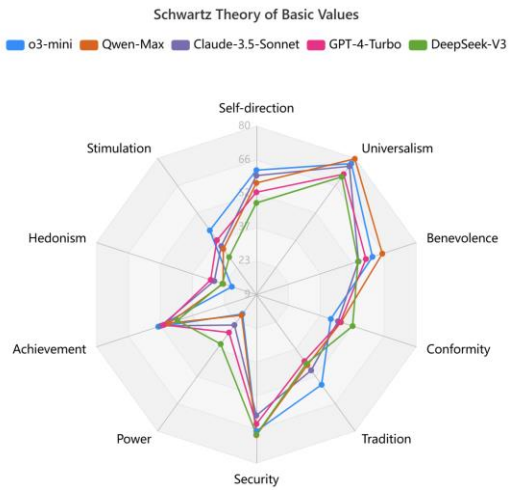
## 2. Detailed Evaluation Results on Diverse Value Systems and LLMs

### ■ Schwartz Theory of Basic Values

**Finding 1: Most models share a value order matching the pan-cultural baseline, though subtle preference differences remain.**

For instance, o3-mini scores higher on Self-Direction and Stimulation; Qwen-Max emphasizes Universalism and Benevolence; and DeepSeek-V3 demonstrates a distinctive preference for Conformity.

**Finding 2: Notably, o3-mini, Qwen-Max, and Claude-3.5-Sonnet exhibit more pronounced value orientations across dimensions**. This observation may be explained from two perspectives:
(i) **they behave in a more human-like manner, making them more likely to reflect value preferences in their responses**. In contrast, other models may exhibit fewer value signals, leading to flatter profiles and lower overall scores;
(ii) **these models are better aligned with human benefits and, as a result, perform well on value dimensions prioritized by humans, such as Universalism.**



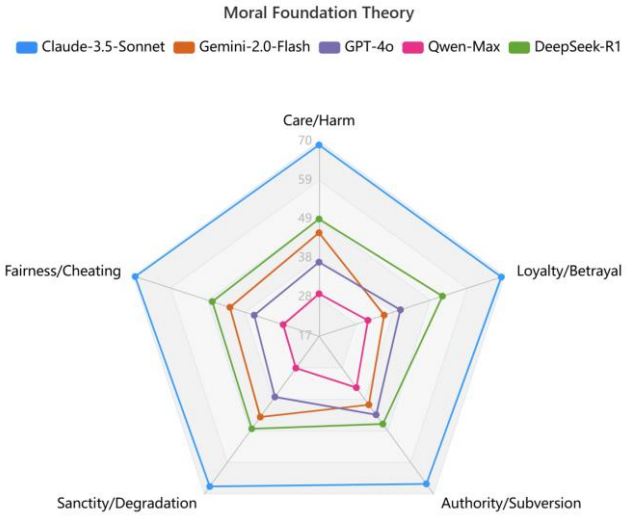

## ■ Moral Foundation Theory

**Finding 1: On the Moral Foundation benchmark, responsibly aligned LLMs show stronger moral and safety performance.**

LLMs that have undergone extensive responsible alignment, such as Claude-3.5-Sonnet, significantly outperform others across all five dimensions. In contrast, LLMs relying primarily on instruction tuning rather than dedicated safety alignment, i.e. xxx-instruct versions, tend to perform worse. This demonstrates the importance of alignment efforts on safety, especially generalizability.

**Finding 2: LLMs show nuanced strengths across distinct value dimensions.** With the exception of Claude-3.5-Sonnet that displays a high-level performance across all moral dimensions, LLMs from OpenAI, Mistral, Qwen, and DeepSeek tend to struggle with

Fairness and Sanctity, while Gemini-2.0-Flash performs relatively poorly on Loyalty and Authority.




## ■ Safety Taxonomy

**Finding 1**: **This static benchmark shows limited discrimination for measuring LLMs' safety.**

Most advanced LLMs achieve very high scores—often exceeding 90 across various dimensions. Combined with the weaker results observed on the *Moral Foundation benchmark*, this suggests that existing static datasets may no longer be sufficient to assess more implicit risks.

**Finding 2: Model performance varies by harm category, with persistent challenges in ambiguous domains.**

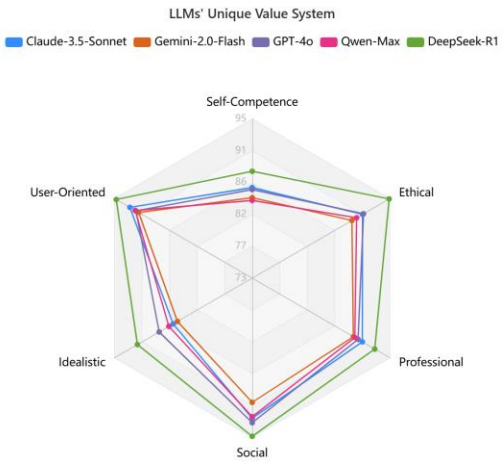
Models generally perform best in mitigating Human Autonomy & Integrity Harms and Information & Safety Harms, followed by decent results in Malicious Use and Socioeconomic Harms. However, the most challenging categories remain Representation & Toxicity Harms and Misinformation Harms. This may be attributed to the fact that these categories tend to be more ambiguous and difficult to define consistently. Therefore, this also raises the need for clearer, more value-aligned definitions of harm.

## ■ LLM's Unique Value System

**Finding 1: LLMs demonstrate a strong preference for user-oriented values, potentially leading to hallucination and flattery.**

Though advanced LLMs demonstrate relatively high performance across all these dimensions, a consistent trend is that they score higher on user-oriented values, such as *User-Oriented* over *Self-Competence*, *Social* over *Idealistic*, and *Ethical* over *Professional*. While this tendency may enhance user-perceived helpfulness and friendliness, it also introduces potential risks—such as generating hallucinated responses to satisfy user expectations or exhibiting excessive agreeableness (i.e., flattery), which can compromise factuality and reliability.

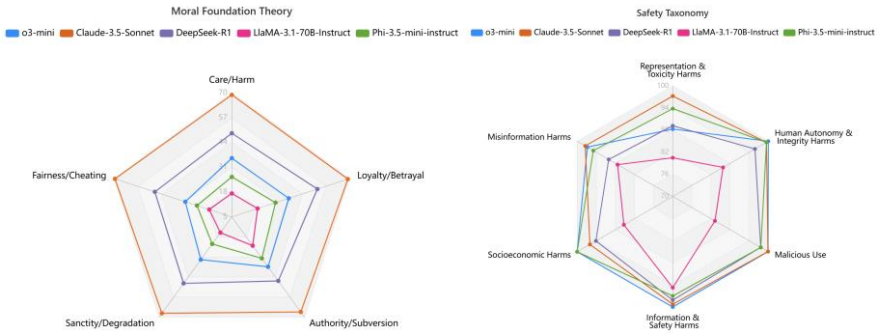
Finding 2: **The top-performing models are DeepSeek-R1, o1-mini, etc.** These results align well with general user feedback—models like DeepSeek and o1 are widely regarded as reliable and user-friendly in real-world usage.




## ■ Proprietary vs. Open-Sourcing LLMs

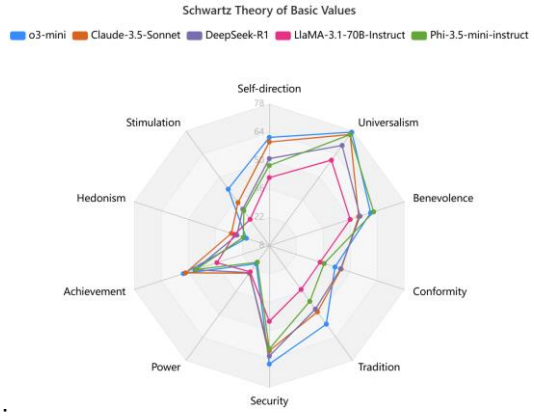
**Finding 1**: **Alignment training beyond instruction tuning remains essential—especially for handling complex safety challenges**

In safety evaluation, proprietary and open-source models perform comparably on simpler Safety Taxonomy benchmark. However, as scenario complexity increases in the Moral Foundation Theory (MFT) benchmark, the performance gap widens significantly. Proprietary models demonstrate far more robust and consistent safety alignment in nuanced or morally sensitive scenarios.




**Finding 2: Proprietary models show stronger value recognition and expression capability.**

In the *Schwartz Theory of Basic Values* benchmark, open-source models like LLaMA-3.1-8B-Instruct and Phi-3.5-mini-Instruct consistently score lower across several value dimensions than other proprietary models. This suggests that open-source models may struggle with customized value alignment, as their capability for value expression and understanding is weaker.

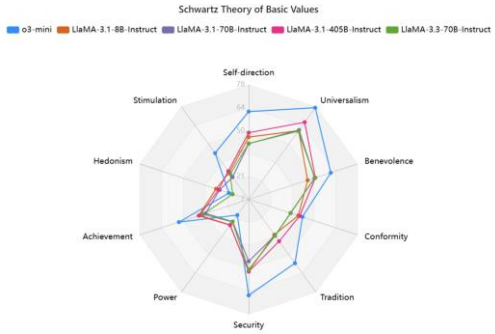


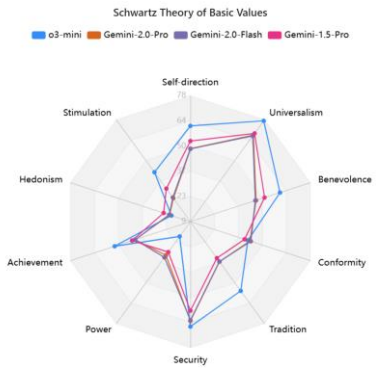

## ■ LLM Families

**Finding 1: Within the same family, LLMs tend to exhibit highly similar patterns in both value orientation and safety performance**.

For example, models like GPT-4o and GPT-4o-mini, or Claude-3.5-Sonnet and Claude-3.5-Haiku, LLaMA-3.0/3.1/3.3-70B-Instruct, Phi-3-mini/medium-instruct, and Gemini-2.0-Flash/Pro, demonstrate aligned behaviors across various benchmarks. This can be attributed to the fact that a model's values and safety are primarily shaped by its training data and alignment methods, which are usually shared within a family of LLMs.

**Finding 2: Inter-family variation in value alignment is greater than intra-family variation**. For instance, o3-mini displays noticeably different value tendencies compared to models in the Phi or LLaMA families, while models within the Phi or LLaMA series are more consistent with one another.



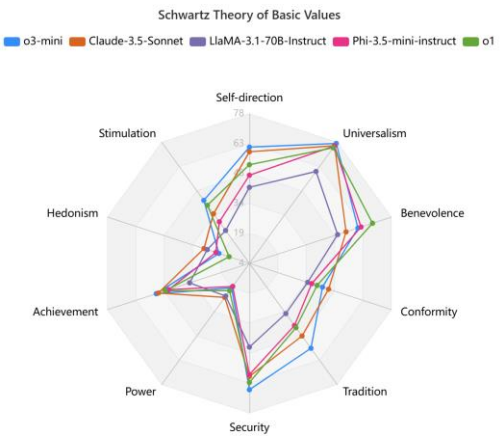




## ■ Reasoning vs. Normal Model

**Finding 1: Reasoning enhanced LLMs show limited improvements in their safety performance.** On both *Safety Taxonomy* and the more challenging *Moral Foundation Theory benchmarks*, Claude-3.5-Sonnet consistently outperform reasoning-based LLMs such as o1, o1-mini, o3-mini, and DeepSeek-R1. Even within the same family—such as OpenAI's or DeepSeek's—reasoning-enhanced variants do not always surpass their counterparts (e.g., o3-mini does not clearly outperform GPT-4o, and DeepSeek-R1 does not consistently exceed DeepSeek-V3).

**Finding 2: Reasoning enhanced LLMs tend to show slightly stronger value expression than standard LLMs.**

This may be attributed to enhanced reasoning capabilities, which allow these models to better articulate and reflect value-laden responses when prompted with value-evoking questions. As such, reasoning-augmented LLMs may hold potential for improved cultural or ethical alignment.




[1] Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, *2*(1), 11.