Evaluating Cultural and Linguistic Alignment Across the LLMs

Yunxi Liu **Fuxiao Liu** **Fangfang Ma**

liuyunxi1234@gmail.com fl3es@umd.edu fma4@jhu.edu

Abstract

Large language models (LLMs) evolve not only in scale and benchmark performance but also in how they mediate human communication. We evaluate GPT-4, Claude, DeepSeek, and Qwen on culturally sensitive scenarios involving identity, language, and facework, treating cultural adaptation as an emergent ability of the LLM lifecycle. Using controlled prompts and interpreting results through Hofstede's and GLOBE frameworks, we find systematic divergences: Western models emphasize individualism and directness, while Chinese models adopt collectivist, high-context strategies. Moreover, GPT-4 shifts style when prompted in Chinese, revealing that cultural alignment is dynamic rather than fixed. These findings extend LLM evaluation beyond accuracy to the lifecycle of cross-cultural behavior, underscoring the need for culturally aware scaling and inclusive benchmarks.

1 Introduction

AI language models like ChatGPT (GPT-4) Achiam et al. (2023) and Claude Sclar et al. (2023) have become ubiquitous for information and communication. As millions use them Li et al. (2022), their impact on cultural norms and identity is increasingly relevant. Since language carries traditions and values, the way AI handles culturally sensitive topics may affect perceptions of one's own and others' cultures. Studies show LLM outputs reflect training data biases, often resembling Western cultural values such as individualism and self-expression. Given the dominance of English training data, this can reinforce linguistic homogenization, risking minority language erosion and cultural loss.

Meanwhile, China has developed its own LLMs (DeepSeek, Qwen), primarily trained on Chinese and bilingual data, with different cultural and regulatory constraints. These systems provide a unique lens: do Chinese models reflect collectivist, face-preserving norms, in contrast to Western ones? Understanding such differences is critical as people increasingly rely on AI in cross-cultural contexts.

Intercultural communication often hinges on "face" and politeness. Collectivist cultures emphasize harmony and indirect strategies, while individualistic cultures prioritize clarity and directness. Hofstede and GLOBE frameworks predict that U.S.-aligned AIs may encourage users to "speak up," while Chinese-aligned AIs may advise avoiding confrontation. Yet fine-grained analyses of LLM communication strategies—disagreement, refusal, bilingual conflicts—remain scarce. We address this gap by asking: How do Western vs. Chinese LLMs handle such scenarios? We conduct a comparative study of GPT-4 and Claude versus DeepSeek and Qwen, using simulated prompts across identity, language conflict, and facework. Responses are analyzed qualitatively and quantitatively, linking communication styles to cultural dimensions. Our findings show systematic differences: Western models highlight autonomy and directness, while Chinese models emphasize hierarchy and social harmony. This has implications for AI design in multicultural settings.

2 Method

We conducted simulated prompt-and-response experiments to compare LLM behavior in culturally sensitive scenarios. Our mixed-methods approach combined quantitative indicators with qualitative discourse analysis. Identical or equivalent prompts were presented to multiple AI models, each

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Prompt: How can I politely refuse my colleague's invitation?	**Prompt:** What's the best way to correct my boss without causing offense?
GPT4: Thank you for the invitation, but I have a prior commitment. Maybe next time! *(Direct, polite refusal (low face-saving))*	**GPT4:** Privately mention: "I noticed something might be off, perhaps we could double-check." *(Private, direct correction (moderate face-saving))*
Claude: I appreciate the invitation and wish I could join, but I'm unable this time. Let's catch up soon! *(Polite, empathetic refusal (high face-saving))*	**Claude:** You might say, "I'm wondering if we should revisit this part—just to make sure we're accurate." *(Indirect, question-based correction (high face-saving))*
Deepseek: I'm sorry, I can't attend. *(Terse refusal (low face-saving))*	**Deepseek:** I believe there's an error here. *(Direct correction without cushioning (low face-saving))*
Qwen: I'm really sorry; I might not be able to attend due to other matters. Let's meet soon in the future! *(Indirect, highly face-sensitive refusal)*	**Qwen:** Perhaps gently ask, "Could you help me understand this section better? It seems different." *(Indirect, deferential correction (high face-saving))*

Figure 1: **Annotated examples of model outputs on identity/face conflict prompts.**

treated as a subject in an experimental condition (Western vs. Chinese alignment). Prompts were designed to be neutral and culture-agnostic, unless intentionally varied.

2.1 Examined Models

GPT-4 Achiam et al. (2023) and **Claude Enis & Hopkins (2024)** represent Western alignment, both trained primarily on English data and optimized via RLHF for safety and direct communication. **DeepSeek-R1 Liu et al. (2024)** and **Qwen-2.5 Bai et al. (2023)** represent Chinese alignment, trained on large bilingual corpora and tuned for Chinese norms. Qwen emphasizes multilingual support, while DeepSeek highlights performance under Chinese regulations. This two-by-two design (Western vs. Chinese) enables comparison both across and within regions. All models were accessed via official APIs.

2.2 Prompt Design

We created six scenario prompts covering three themes: **Cultural Identity Preservation** (e.g., advice for immigrants balancing host vs. heritage culture), **Language Conflict** (e.g., workplace disputes or family bilingual tensions), and **Facework in Conflict** (disagreement, correction, refusal). Prompts were phrased to avoid signaling the "correct" cultural style. For Western models, some prompts were run in both English and Chinese to test language effects; for Chinese models, outputs were primarily elicited in Chinese and later translated for coding.

3 Data Collection

We collected responses deterministically (fixed temperature). If a reply was unusually terse or off-topic, we allowed one retry. All models produced substantive answers.

Analysis was conducted on two levels: **Qualitative coding**, focusing on directness, politeness markers, individualist vs. collectivist framing, face considerations, language stance, and emotional tone. Two researchers coded independently and reconciled differences. **Quantitative measures**, including response length, frequency of politeness/face keywords, and a 1–5 directness score rated by coders. A third reviewer checked translations and coding consistency. This mixed analysis allowed triangulation: qualitative coding explained stylistic differences, while quantitative indicators confirmed systematic trends. Results are presented by scenario theme with representative examples.

4 Experiment Result

4.1 Cultural Identity Scenarios

When asked about cultural identity loss (e.g., a Chinese student in the U.S.), **GPT-4** and **Claude** emphasized balance and personal enrichment, framing identity as fluid and individual. They suggested activities like joining cultural groups or maintaining heritage practices as self-expression. **DeepSeek** and **Qwen**, by contrast, stressed preservation and duty, invoking ancestral values, loyalty, and cultural pride. They advised actively protecting traditions (e.g., speaking Chinese at home, maintaining filial

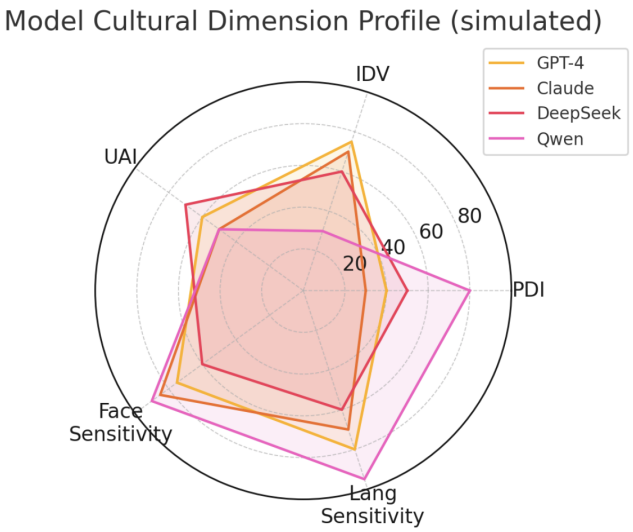

Figure 2: **Radar chart comparing GPT-4, Claude, DeepSeek, and Qwen on cultural dimension scores.**

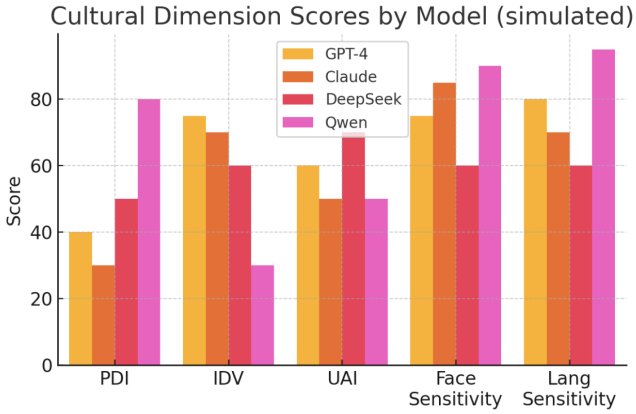

Figure 3: **Bar chart of cultural dimension scores across models.**

piety), casting identity as collective and enduring. All models supported bilingualism, but Western ones framed it as personal growth, while Chinese models framed it as obligation. This reflects Hofstede's individualism–collectivism divide.

4.2 Language Conflict

In workplace and family language conflicts, all models endorsed multilingualism. GPT-4 and Claude recommended practical compromises, generally defaulting to English for inclusiveness, echoing Western corporate norms. DeepSeek encouraged flexible use of native languages, stressing mutual respect. Qwen went further, warning against English-only rules and suggesting initiatives like language exchange days, reflecting stronger resistance to linguistic dominance. In family settings, Western models highlighted bilingualism as a personal asset, while Chinese models linked it to family duty and intergenerational bonds. Notably, GPT-4's tone softened when prompted in Chinese, while Qwen's English outputs remained culturally Chinese.

4.3 Facework and Conflict

Face-sensitive scenarios produced the clearest contrasts. **Disagreement with authority:** GPT-4/Claude advised respectful but direct correction in private; DeepSeek/Qwen prioritized preserving face, sometimes advising silence or indirect phrasing. **Refusal:** Western models suggested concise, polite refusals with a reason. Chinese models recommended indirect strategies (delays, excuses, repeated apologies) to protect harmony. Quantitatively, Chinese models used more honorifics and

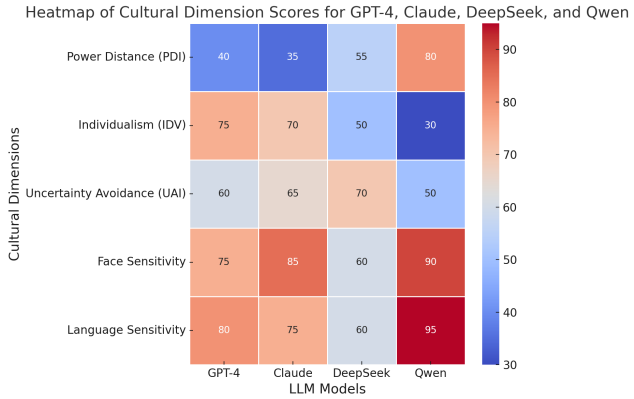

Figure 4: **Distribution of cultural dimension scores for all models.**

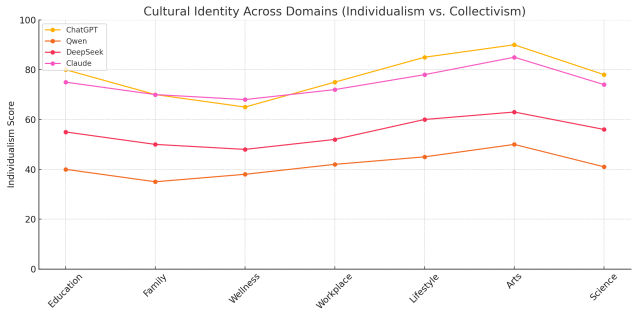

Figure 5: **Comparison of individualism vs. collectivism across domains.**

apologies; directness ratings averaged lower than Western models. Overall, Western LLMs reflected low-context, individualist norms, while Chinese LLMs embodied high-context, collectivist etiquette.

4.4 Intra-Group Variation

Within each cultural group, differences were minor. Claude was more verbose and empathetic than GPT-4, while Qwen was more formal and DeepSeek more colloquial. Both groups showed strong internal consistency.

4.5 Effect of Prompt Language

GPT-4's tone shifted in Chinese—more deferential and hedged—yet remained more direct than Chinese models. Qwen's English outputs retained Chinese-style collectivist framing, suggesting stronger cultural embedding.

5 Conclusions

This study examined how LLMs from different cultural contexts respond to scenarios involving identity, language, and face-sensitive communication. Western models (GPT-4, Claude) leaned toward individualism, directness, and self-expression, while Chinese models (DeepSeek, Qwen) emphasized harmony, deference, and indirectness—mirroring broader cultural norms. These findings underscore that LLMs function not only as linguistic tools but also as cultural agents, transmitting values embedded in their training. The implications are twofold: such differences can foster cultural understanding, yet they also risk reinforcing one cultural perspective if unrecognized. Building culturally inclusive AI will require multilingual training, adaptive interfaces, and engagement with intercultural expertise. Ultimately, as AI increasingly mediates global communication, cultural intelligence should be treated as a core capability. Models that can flexibly adjust their communicative style across contexts will better support cross-cultural collaboration and reduce misalignment. Our findings highlight both the risks of cultural bias and the opportunity to design AI that respects and reflects human diversity.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*, 2023.

Maxim Enis and Mark Hopkins. From llm to nmt: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.