Somatic in the East, Psychological in the West?: Investigating Clinically-Grounded Cross-Cultural Depression Symptom Expression in LLMs

Anonymous ACL submission

Abstract

Prior clinical psychology research shows that Western individuals with depression tend to report psychological symptoms, while Eastern individuals report somatic ones. We test whether Large Language Models (LLMs), which are increasingly used in mental health, reproduce these cultural patterns by prompting them with Western or Eastern personas. Results show that LLMs largely fail to replicate the patterns when prompted in English, though prompting in major Eastern languages (i.e., Chinese, Japanese, and Hindi) improves alignment in several configurations. Our analysis pinpoints two key reasons for this failure: the models' low sensitivity to cultural personas and a strong, culturally invariant symptom hierarchy that overrides cultural cues. These findings reveal that while prompt language is important, current general-purpose LLMs lack the robust, culture-aware capabilities essential for safe and effective mental health applications.

1 Introduction

Large Language Models (LLMs) are becoming central to mental health applications. They now support a wider range of uses, including conversational agents in online therapy, synthetic data generation for training and research, and automated classification of mental health conditions (Hua et al., 2024; Obradovich et al., 2024; Lawrence et al., 2024; Stade et al., 2024).

Culture plays a significant role in how individuals express mental health concerns. Cultural differences have been observed in how people report symptoms on questionnaires (Parker et al., 2001; Arnault et al., 2006; Ryder et al., 2008), express symptoms in clinical interviews (Ryder et al., 2008; Biswas et al., 2016), and choose language and topics on social media and online peer support communities (De Choudhury et al., 2017; Loveys et al., 2018; Rai et al., 2024).

Given the high-stakes nature of mental health applications, measuring the cultural alignment of LLMs against established clinical findings is critical for developing culturally safe and sensitive tools. If LLMs disproportionately reflect certain cultural perspectives while neglecting others, they risk reinforcing disparities in mental health care (Shah et al., 2019). For example, depression detection models built on LLMs may fail to identify individuals from non-Western cultures who express depression differently (Ali et al., 2024).

Despite growing interest in LLM alignment, few studies have examined cultural alignment in mental health contexts, leaving a critical gap in our understanding of their reliability across diverse populations. This study addresses this gap by testing whether LLMs reflect clinically observed culturally specific symptom patterns by comparing their responses across Western and Eastern cultural contexts. Clinical psychology literature highlights a well-established cultural distinction: individuals from Western cultures are more likely to report psychological symptoms, while those from Eastern cultures tend to report somatic symptoms (Kleinman, 1982; Tsoi, 1985; Ryder et al., 2008; Arnault et al., 2006; Parker et al., 2005; Biswas et al., 2016; Dere et al., 2013; Kirmayer and Ryder, 2016). We investigate how this cultural distinction is reflected in LLMs' responses. For example, when prompted with a description of an American or Japanese person with depression, do American personas elicit more psychological symptoms, and Japanese personas more somatic ones?

2 Background

2.1 LLMs applications in Mental Health Contexts

LLMs are increasingly used in mental health. A recent review shows a surge in related publications in 2023 (Hua et al., 2024). Their applications fall into three broad areas (Hua et al., 2024): 1) conversational agents for digital companionship and emotional support (Fu et al., 2023; Lai et al., 2023; Ma et al., 2024; Yao et al., 2023; Lee et al., 2024; Zhang et al., 2023; Kumar et al., 2022), 2) resource enrichment, such as generating synthetic data and educational materials (Yang et al., 2023; Kumar et al., 2023; Yang et al., 2024), and 3) classification tasks for conditions like depression severity and suicide risk (Yang et al., 2023, 2024; Xu et al., 2024; Lamichhane, 2023; Qi et al., 2024; Nguyen and Pham, 2024). These use cases demonstrate the expanding role of LLMs in mental health support, content development, and diagnostic modeling.

2.2 Cultural Differences in Depression Symptom Expressions

In clinical psychology studies, numerous studies support that individuals from Western cultures tend to emphasize psychological symptoms, while those from Eastern cultures tend to emphasize somatic symptoms (Kleinman, 1982; Tsoi, 1985; Ryder et al., 2008; Arnault et al., 2006; Parker et al., 2005; Juckett and Rudolph-Watson, 2010; Biswas et al., 2016; Dere et al., 2013; Kirmayer and Ryder, 2016).

For instance, a series of studies (Ryder et al., 2008; Dere et al., 2013) found that Chinese patients consistently reported more somatic symptoms in interviews and problem reports compared to their Euro-Canadian counterparts, who emphasized psychological symptoms. Another study focused on depression in Japanese and American college women, finding that Japanese participants reported higher overall somatic distress. (Arnault et al., 2006). Similarly, Parker et al. found that 60% of Malaysian Chinese patients presented with somatic symptoms of depression, compared to only 13% of Australian Caucasians. While Chinese patients scored higher on somatic items in an inventory, they were less likely to acknowledge psychological symptoms (Parker et al., 2001).

Diagnostic practices also reflect this divide, with Indian psychiatrists prioritizing somatic symptoms (e.g., pain, sleep issues) and American psychiatrists prioritizing cognitive and emotional ones (e.g., pessimism about the future) (Biswas et al., 2016).

Together, these studies highlight cultural differences in how depression is expressed: Eastern populations tend to show somatic symptoms, while Western populations emphasize psychological ones. The most widely accepted explanation for somatization among Eastern populations is that it offers a socially safer way to express mental health problems in cultures where mental illness is highly stigmatized. By framing distress in physical terms, individuals can seek support without being labeled mentally ill (Link et al., 1997; Goldberg and Bridges, 1988; Barney et al., 2006; Kung and Lu, 2008; Juckett and Rudolph-Watson, 2010).

2.3 Bias in LLMs

2.3.1 Cultural Bias in LLMs

Several studies have demonstrated that cultural bias exists across different models, typically using the ersonas method to quantify cultural bias in LLMs. (Santy et al., 2023; Cao et al., 2023; AlKhamissi et al., 2024; Kharchenko et al., 2024; Rao et al., 2024).

It is well known that the cultural values reflected in LLMs tend to align more closely with American values or values of English-speaking countries (Johnson et al., 2022; Santy et al., 2023; Cao et al., 2023; AlKhamissi et al., 2024; Rao et al., 2024). Prompting LLMs in a country's local language has been shown to improve cultural alignment (Lin et al., 2022; Cao et al., 2023; AlKhamissi et al., 2024).

These studies often use sociological benchmarks such as the World Values Survey and Hofstede's cultural dimensions to assess cultural alignment. While these benchmarks are valuable for understanding general cultural values and dimensions, they do not directly capture cultural bias specifically within the nuanced domain of mental health symptom expression.

2.3.2 Bias in LLMs in Mental Health Contexts

Several studies have examined biases in LLMs within mental health contexts. One study focusing on Borderline Personality Disorder (BPD) and Narcissistic Personality Disorder (NPD) found that GPT-3.5 and GPT-4 exhibited gender bias in diagnostic assessments, particularly against women (Chansiri et al., 2024). Another study investigated classification performance of 10 different LLMs across various demographic factors. While models generally performed well with respect to gender and age, their performance varied when factors such as religion and nationality were considered (Wang et al., 2024).

These findings highlight the importance of evaluating LLM bias in mental health applications, given the growing use of LLMs in this domain and the scarcity of research addressing cultural bias in such contexts.

3 Research Hypotheses

Building upon the background reviewed in §2, we investigate whether LLMs select depression symptoms in ways consistent with cultural patterns identified in clinical psychology.

> **H1.** LLMs select psychological symptoms more often for Western cultural personas, and somatic symptoms more often for Eastern cultural personas.

> **H2.** Prompts written in the local language of a country increase cultural alignment in symptom selection.

4 Task Design for Hypothesis Testing

We assign the model a cultural persona with depression (e.g., *an American person with depression*) and prompt it to select symptoms from a predefined list of 14 depression symptoms (see Table 3), which are extracted from the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) (Association et al., 2013). Following prior clinical research (Ryder et al., 2008), these symptoms are categorized as either somatic (e.g., fatigue, sleep disturbance) or psychological (e.g., depressed mood, worthlessness) symptoms.

We use two prompt forms: an implicit culture prompt (ICP) providing only general cultural context, and an explicit culture prompt (ECP) that directly instructs the model to consider the persona's cultural background, allowing us to assess the effect of explicit instruction. Full prompts are provided in Appendix §A.3.

While we refer to cultural personas, we operationalize them at the country level, using national identity as a proxy for broader cultural context. Specifically, we use Japanese, Chinese, and Indian to represent Eastern cultural groups, and the American, Canadian, and Australian to represent Western cultural groups. This follows conventions in clinical psychology research (Parker et al., 2001; Arnault et al., 2006; Ryder et al., 2008; Biswas et al., 2016), where cultural groupings are often defined nationally. We acknowledge that this simplification may overlook within-region and -country variation, but we adopt it for comparability with prior clinical research and to maintain experimental clarity.

Formally, let S_{som} and S_{psy} denote the sets of somatic and psychological symptoms. Let $\mathcal{C}_W$ and $\mathcal{C}_E$ represent the set of Western and Eastern cultural groups, and let $x \in \{\mathrm{I}, \mathrm{E}\}$ indicate the prompt type, either implicit (ICP) or explicit (ECP) culture prompt. Then, for each cultural group c, we denote the probability that the model selects symptom s as $P(s|\mathrm{P}_x^c)$ where P_x^c is the prompt corresponding to cultural persona c under prompt type x.

With $g \in \{\text{Western}, \text{Eastern}\}$, we compute the group-level selection probability for each symptom as:

$$P(s \mid \mathrm{P}_x^g) = \frac{1}{|\mathcal{C}_g|} \sum_{c \in \mathcal{C}_g} P(s \mid \mathrm{P}_x^c)$$

The average selection probability for symptom types is then:

$$P(\text{somatic} \mid \mathrm{P}_x^g) = \frac{1}{|\mathcal{S}_{\text{som}}|} \sum_{s \in S_{\text{som}}} P(s \mid \mathrm{P}_x^g)$$

$$P(\text{psychological} \mid \mathrm{P}_x^g) = \frac{1}{|\mathcal{S}_{\text{psy}}|} \sum_{s \in S_{\text{psy}}} P(s \mid \mathrm{P}_x^g)$$

We define the cultural alignment as the difference in somatic selection:

$${}_1\mathcal{A}_x = P(\text{somatic} \mid \mathrm{P}_x^{\text{Eastern}}) - P(\text{somatic} \mid \mathrm{P}_x^{\text{Western}})$$

, which is equivalent to $P(\text{psychological} \mid \mathrm{P}_x^{\text{Western}}) - P(\text{psychological} \mid \mathrm{P}_x^{\text{Eastern}})$, given that symptoms are exhaustively categorized as somatic or psychological. A positive ${}_1\mathcal{A}_x$ indicates that models follow clinical findings by selecting more somatic symptoms for Eastern personas and more psychological symptoms for Western personas.

Following recent studies that prompts written in non-English languages can elicit responses that are more culturally aligned with the culture of the language (Lin et al., 2022; Cao et al., 2023; AlKhamissi et al., 2024), we test prompts written in the local language spoken in each Eastern cultural group. Specifically, we use Japanese for Japanese, Chinese for Chinese, and Hindi for Indian. While we acknowledge that multiple languages and dialects are spoken within each country, we safely selected the most widely spoken language in each country. For simplicity, we refer to English prompts as the English Language Prompt (ENG-P), and those written in local language as the Local Language Prompt (LOC-P). Let $l(c)$ be a function that maps each Eastern cultural group $c \in \mathcal{C}_{\text{Eastern}}$ to its local language (e.g., $l(\text{Japan}) = \text{Japanese}$).We denote ${}_1\mathcal{A}_x(l(c))$ as the cultural alignment tested by LOC-P for Eastern cultural group c, and ${}_1\mathcal{A}_x(Eng)$ as that by ENG-P.

$$
\begin{aligned}
{}_1\mathcal{A}_x(l(c)) = P(\text{somatic} \mid \mathrm{P}_x^{\text{Eastern}}, l(c)) \\
- P(\text{somatic} \mid \mathrm{P}_x^{\text{Western}}, \text{English})
\end{aligned}
$$

This formulation allows us to test the hypotheses as:

H1 is supported when ${}_1\mathcal{A}_x > 0$, and
H2 is supported when ${}_1\mathcal{A}_x(l(c)) > {}_1\mathcal{A}_x(Eng)$.

Each model is tested across three symptom-choice conditions (LLMs can select one, three, or five symptoms per each iteration), two prompt types (I/E), and six countries (three Western, three Eastern), yielding 36 configurations under ENG-P and 18 under LOC-P. We run 100 iterations per setting to ensure reliable results.

Throughout the paper, we use both symbolic notation and plain language interchangeably to improve readability and ease of understanding.

4.1 Five Language Models for Evaluation

We tested five LLMs in total. Four open-source models were selected for their accessibility and replicability: Llama-3.1-8B-it (Touvron et al., 2023), Gemma-7B-it (Team et al., 2024), Qwen-2.5-7B-it (Bai et al., 2023), and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025). We additionally included GPT-4o (OpenAI et al., 2024), as it is the most widely used. In our experiments, we tested temperatures of 0.5, 0.7, and 1.0, but observed no notable differences in output. We report results generated with temperature 0.7, a commonly used default value in prior studies (Chansiri et al., 2024; AlKhamissi et al., 2024).

5 Results

5.1 Cultural Alignment in English Prompts (H1)

Figure 1(a) shows each model's alignment level under the ENG-P condition, for both P_I and P_E. The x-axis presents cultural alignment ${}_1\mathcal{A}$, defined as $P(\text{somatic} \mid \mathrm{P}_x^{\text{Eastern}}) - P(\text{somatic} \mid \mathrm{P}_x^{\text{Western}})$. Positive values of ${}_1\mathcal{A}$ indicate alignment with prior clinical psychology findings that somatic symptoms are more common in Eastern contexts. The y-axis shows $P(\text{somatic} \mid \mathrm{P}^{\text{Eastern}})$, the average proportion of S_{som} selected by Eastern personas. Since 5 out of 14 symptoms are somatic, the random baseline is 5/14 (≈0.357); values above this indicate a bias toward somatic symptom selection. We use the y-axis to examine whether higher absolute somatic bias corresponds to larger cultural gaps ${}_1\mathcal{A}$.

Overall, LLMs' behaviors do not align with expectations under the ENG-P condition. Of the 30 tested settings (5 models $\times$ 3 choice conditions $\times$ 2 prompt types), only three—Llama (one-choice, P_E), GPT (one-choice, P_E), and Qwen (one-choice, P_I)—show patterns consistent with prior clinical findings. In general, absolute values of ${}_1\mathcal{A}$ is small. Differences are particularly marginal in the three- and five-choice conditions, where absolute values of ${}_1\mathcal{A}$ typically range from 0.03 to 0.05. This suggests that assigning Western or Eastern cultural personas has limited influence on LLM symptom selection in multi-choice conditions. A full breakdown of symptom selection rates across all model and prompt configurations is available in Table 6 in §A.5 of the Appendix[1].

Llama, GPT, and Qwen exhibit a somatic symptom selection bias in the one-choice condition, primarily driven by a strong preference for s2 (Decreased energy, tiredness, and fatigue; Somatic). In contrast, DeepSeek shows a psychological symptom selection bias, largely due to a strong preference for s3 (Depressed mood; Psychological) (See §A.4 in the Appendix). However, higher absolute somatic bias, captured by $P(\text{somatic} \mid \mathrm{P}^{\text{Eastern}})$, does not correspond to larger cultural gaps in alignment (${}_1\mathcal{A}$).

In summary, *H1 is not supported*, as the majority of experimental conditions fail to align with findings from prior clinical psychology research.

5.2 Effect of Language on Alignment (H2)

Interestingly, of the 30 settings, 15 exhibit increased alignment under the LOC-P (Figure 1(b)). Positive values of ${}_1\mathcal{A}_x(l(c)) - {}_1\mathcal{A}_x(Eng)$ on the x-axis indicates the alignment increase. Llama shows improvement in the three- and five-choice conditions, Qwen in the five-choice condition, and Gemma primarily in the one-choice condition. DeepSeek demonstrates consistent alignment across all conditions, whereas GPT shows decreased alignment throughout. These results suggest that prompt language can affect alignment levels in some models. Detailed results are provided

[1]While we aim to make the main text self-contained, the number and complexity of experimental conditions make it impractical to include all results. To support transparency and reproducibility, we provide detailed model-wise results and statistical comparisons in the Appendix.

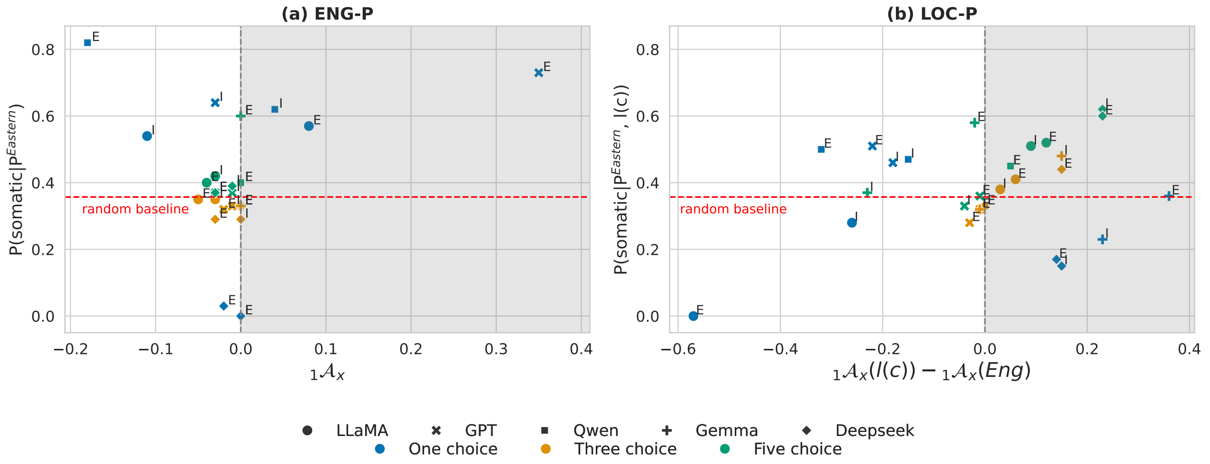


Figure 1: In (a), the x-axis shows cultural alignment $_1\mathcal{A}_x$ under ENG-P condition; values greater than 0 indicate alignment with prior clinical psychology findings. In (b), values greater than 0 on the x-axis indicate an *increase* in alignment under the LOC-P condition. Higher values on the y-axis reflect a stronger tendency for the model to choose S_{som}. I and E indicate implicit (ICP, P_I) and explicit cultural prompt (ECP, P_E), respectively. Culturally aligned regions are shaded to help readers visually identify expected model behavior.

in Table 7 and 8 in §A.5 of the Appendix.

To assess the impact of language on alignment, we conduct paired t-tests, as in Table 1. Each t-test evaluates alignment within a distinct experimental condition (e.g., model, choice condition, or prompt type). As our goal is not to test a single global hypothesis, but rather to probe how alignment shifts across various independent conditions, we do not apply multiple comparisons correction. Overall, alignment increases slightly (t=-0.13), though the effect is not statistically significant (p=0.90). At the model level, both DeepSeek and Gemma show improved alignment, with statistical significance observed only for DeepSeek (p<0.0005). By choice condition, both the three- and five-choice conditions show increases, with a larger effect in the three-choice; however, neither reached statistical significance (p=0.07 and 0.30, respectively). At the prompt level, P_I also shows a modest, non-significant increase (p=0.74). These results suggest that not all models are equally sensitive to language-induced cultural cues, and the effectiveness of prompt language depends on the choice condition and prompt type. Notably, only DeepSeek shows a statistically significant improvement, suggesting the overall effect of prompt language remains limited.

We note that higher absolute somatic bias, captured by $P(\text{somatic} \mid \text{P}^{\text{Eastern}})$, does not correspond to the alignment level changes $_1\mathcal{A}_x(l(c)) - {_1\mathcal{A}_x}(Eng)$.

In summary, the *results do not support H2* overall, though DeepSeek shows a statistically significant increase in alignment, providing model-level support.

5.3 Individual Symptom Level Analysis

We also conducted a more detailed, symptom-level analysis of psychological symptoms. While prior research generally suggests Eastern somatization and Western psycholization, not all psychological symptoms have been consistently reported as more prevalent among Western populations or showed statistically significant differences between Western and Eastern samples (Parker et al., 2001; Biswas et al., 2016; Dere et al., 2013). We first identified psychological symptoms with statistically significant differences between Western and Eastern populations ($p < 0.05$). These were drawn from prior studies that conducted symptom-level analyses (Parker et al., 2001; Biswas et al., 2016; Dere et al., 2013). We then matched these symptoms with their corresponding entries in the DSM-5. Symptoms not included in the DSM-5 were excluded from this analysis. Finally, we examined whether LLMs exhibited similar patterns of statistically significant differences ($p < 0.05$) in symptom selection across Eastern and Western cultural personas. Please refer to §7 for a discussion of why we did not perform individual-level analyses for somatic symptoms.

Australian-Chinese Pair. A study by Parker et al. (2001) found that *Depressed mood* and *Loss of interest* were significantly more frequently re-

	All	Llama	GPT	Qwen	Gemma	DeepSeek	One Choice	Three Choice	Five Choice	P_I	P_E
t-stat	-0.13	0.79	2.15	1.03	-0.93	**-10.02**	0.89	-2.10	-1.09	-0.34	0.08
p-value	0.90	0.46	0.09	0.35	0.39	**0.00**	0.39	0.07	0.30	0.74	0.94

Table 1: Paired t-test by model, experiment condition, and prompt type. Statistically significant values ($p < 0.05$) are bolded.

ported by Australian patients, whereas *Suicidal thoughts* were more commonly reported by Chinese patients. In our results, alignment with Parker et al. (2001) is limited: for *Depressed mood*, only one condition (GPT-LOC-P, One choice) shows alignment. For *Loss of interest*, only the (GPT, LOC-P, Three choice) condition aligns with the clinical findings. *Suicidal thoughts* has not alignment condition.

Canadian-Chinese Pair. A study by Dere et al. (2013) reported that *Depressed mood* was significantly more likely to be expressed by Chinese patients than Canadian patients. We find alignment with this result in three conditions: (Deepseek, ENG-P, One choice), (LLaMA, LOC-P, One choice), and (GPT, LOC-P, Three choice).

American-Indian Pair. A study by Biswas et al. (2016) found that American psychiatrists placed greater emphasis on *Decreased interest in pleasurable activities*, *Pessimistic view of the future*, and *Ideas of self-harm or suicide*. In our results, *Decreased interest in pleasurable activities* aligns in only one condition: (GPT, LOC-P, One choice). For *Pessimistic view of the future*, two conditions show alignment: (Deepseek, ENG-P, One choice) and (LLaMA, ENG-P, Five choice). *Ideas of self-harm or suicide* has no alignment condition.

Overall, across 30 experimental settings (3 choice conditions $\times$ 2 prompt types $\times$ 5 models) for each symptom, only zero to three conditions per symptom showed alignment with prior clinical psychology studies. Alignment at the individual symptom level for psychological symptoms is limited mainly due to the small absolute values of $P(s \mid \mathrm{P}_x^{\text{Eastern}}) - P(s \mid \mathrm{P}_x^{\text{Western}})$.

5.4 Sensitivity to Cultural Personas

To further quantify the models' limited ability to distinguish between Western and Eastern personas in symptom selection, we examine each model's *Persona sensitivity*. Persona sensitivity refers to how well a model differentiates symptom selection patterns between Western and Eastern cultural personas, measured by the cosine similarity between the distributions $P(s \mid \mathrm{P}_x^{\text{Eastern}})$ and $P(s \mid \mathrm{P}_x^{\text{Western}})$ within the same prompt type. Lower cosine similarity values indicate higher sensitivity.

Overall, models exhibit significantly low persona sensitivity under the ENG-P condition, with cosine similarity values ranging from 0.77 to 1.00 (Table 2). Specifically, sensitivity is low in the three- and five-choice settings (e.g., cosine similarity => 0.95), suggesting that cultural distinctions weaken as the number of selectable symptoms increases. For Llama and GPT, persona sensitivity improves from P_I to P_E (e.g., Llama: 0.98 $\rightarrow$ 0.95, GPT: 0.99 $\rightarrow$ 0.77 in one-choice), indicating that explicitly prompting for cultural consideration help these models better distinguish between Western and Eastern cultural personas. However, the effectiveness of P_E remains limited: sensitivity values are still high, and the effect is not observed in Qwen or DeepSeek. Under the LOC-P condition, persona sensitivity increases in 28 out of 30 experimental conditions (3 choice conditions $\times$ 2 prompt types $\times$ 5 models). As with the ENG-P condition, the impact of P_E is confined to specific experimental configurations.

Importantly, low persona sensitivity under the ENG-P condition suggests that internal model tendencies override the influence of cultural personas or prompt variations. Although the findings indicates that prompting in local languages leads to more distinct symptom selection behaviors between Western and Eastern personas, it is important to note that higher sensitivity does not necessarily imply better cultural alignment. For example, while Llama shows greater sensitivity under the LOC-P, this does not translate to improved alignment with clinically observed patterns.

5.5 Symptom Preference Hierarchy

Persona sensitivity analysis reveals that models tend to select similar symptoms for Western and Eastern cultural personas, particularly under the ENG-P condition. To further examine this universal symptom preference, we averaged $P(s|\mathrm{P}_x^c)$ across 180 experimental settings (6 countries $\times$ 3 choice conditions $\times$ 2 prompt types $\times$ 5 models) for ENG-P and 90 experimental settings (3 coun-

	Llama		GPT		Qwen		Gemma		DeepSeek	
	ENG-P	LOC-P	ENG-P	LOC-P	ENG-P	LOC-P	ENG-P	LOC-P	ENG-P	LOC-P
One choice	0.98/0.95	0.31/0.51	0.99/0.77	0.91/0.92	0.99/0.98	0.63/0.83	0.89/0.89	0.84/0.00	0.95/0.99	0.39/0.32
Three choice	0.99/0.98	0.67/0.81	1.00/0.95	0.99/0.99	1.00/1.00	0.96/0.83	1.00/1.00	0.95/0.97	1.00/1.00	0.95/0.94
Five choice	0.98/0.98	0.84/0.85	1.00/1.00	0.96/0.98	1.00/1.00	0.88/0.83	1.00/1.00	0.81/082	1.00/0.99	0.86/0.84

Table 2: Cosine similarities between $P(s \mid \mathtt{P}_x^{\text{Eastern}})$ and $P(s \mid \mathtt{P}_x^{\text{Western}})$ distributions under the ENG-P and LOC-P condition. Two cosine similarity values under each prompt correspond to the consine values for $\mathtt{P}_I$ and $\mathtt{P}_E$.

tries × 3 choice conditions × 2 prompt types × 5 different models) for LOC-P. Before averaging, $P(s|\mathrm{P}_x^c)$ values were normalized by the maximum possible selection rate in each choice condition: 1/3 for three-choice and 1/5 for five-choice. For example, in the three-choice setting (3 choices × 100 iterations = 300 total selections), a symptom chosen in all iterations would have a maximum selection rate of 100/300 = 1/3.

Table 3 shows that under the ENG-P condition, LLMs consistently favor certain symptoms, particularly s2 (Decreased energy, tiredness, and fatigue; Somatic), s3 (Depressed mood; Psychological), and s8 (Loss of interest or motivation; Psychological) with average selection rates of 0.6 for s2 and s3, and 0.4 for s8. In contrast, symptoms like s12 (Suicidal ideas; Psychological), s10 (Poor memory; Psychological), and s7 (Indecisiveness; Psychological) are rarely chosen (0.00–0.02). A similar trend appears under the LOC-P condition (Table 4), where s2 and s3 remain the most selected ($\simeq$ 0.6). While s8 was still frequently selected with 0.27, s1 (Anger and irritability; Psychological) rises to 0.41, becoming the third most selected. s7, s10 and s12 shows slight increases in proportion to 0.06 but remained rarely selected, along with s6 (Inattention; Psychological) and s9 (Pessimism; Psychological).

These results suggest that LLMs exhibit a hierarchical understanding of depression symptoms, consistently identifying some as more representative. This hierarchy likely reflects the frequency of each symptom in training data, clinical descriptions, or public discourse around depression. Importantly, the models' overarching preference patterns often outweigh the influence of cultural personas. The underrepresentation of s12 may also result from safety and moderation mechanisms that suppress suicide-related content (Li et al., 2024), potentially limiting LLM utility in identifying patients with critical symptoms.

Symptom Name	Average	StdDev
s1: Anger and irritability (Psychological)	0.22	0.36
s2: Decreased energy, tiredness, and fatigue (Somatic)	**0.62**	0.37
s3: Depressed mood (Psychological)	**0.59**	0.43
s4: Genitourinary symptoms (Somatic)	0.07	0.23
s5: Hyperactivity and agitation (Somatic)	0.10	0.25
s6: Inattention (Psychological)	0.08	0.22
s7: Indecisiveness (Psychological)	0.02	0.09
s8: Loss of interest or motivation (Psychological)	**0.39**	0.40
s9: Pessimism (Psychological)	0.10	0.22
s10: Poor memory (Psychological)	0.01	0.03
s11: Sleep disturbance (Somatic)	0.15	0.30
s12: Suicidal ideas (Psychological)	0.00	0.01
s13: Weight and appetite change (Somatic)	0.06	0.17
s14: Worthlessness and guilt (Psychological)	0.09	0.20

Table 3: Average and standard deviation of selected symptom proportions under the ENG-P condition.

Symptom Name	Average	StdDev
s1: Anger and irritability (Psychological)	**0.41**	0.42
s2: Decreased energy, tiredness, and fatigue (Somatic)	**0.64**	0.41
s3: Depressed mood (Psychological)	**0.60**	0.43
s4: Genitourinary symptoms (Somatic)	0.14	0.30
s5: Hyperactivity and agitation (Somatic)	0.14	0.30
s6: Inattention (Psychological)	0.06	0.14
s7: Indecisiveness (Psychological)	0.06	0.18
s8: Loss of interest or motivation (Psychological)	0.27	0.39
s9: Pessimism (Psychological)	0.06	0.17
s10: Poor memory (Psychological)	0.04	0.08
s11: Sleep disturbance (Somatic)	0.20	0.30
s12: Suicidal ideas (Psychological)	0.06	0.14
s13: Weight and appetite change (Somatic)	0.11	0.23
s14: Worthlessness and guilt (Psychological)	0.10	0.24

Table 4: Average and standard deviation of selected symptom proportions under the LOC-P condition.

5.6 Robustness Check: The Inverse Task of Cultural Attribution

To further assess the robustness of our findings, we conducted a complementary "cultural attribution" task. In this inverse task, models were given a symptom and asked to choose which of two cultural personas (i.e., one Western and one Eastern) was more likely to express it. The detailed methodology and results are presented in §B in Appendix.

Consistent with our primary findings, the models largely failed to replicate clinically observed patterns in this attribution task as well. The models often showed a consistent bias, favoring one cultural group across all symptom types. This result from the inverse task reinforces our main conclusion: the LLMs we evaluated lack the nuanced, clinically-grounded reasoning required for cross-cultural mental health contexts.

6 Discussion

Overall, our findings reveal substantial inconsistencies between LLM outputs and widely-recognized clinical patterns of depression symptom expression across cultures. Several factors may contribute to this misalignment.

First, the quality and scope of training data are likely key factors. None of the evaluated models are fine-tuned on mental health–specific or clinical psychology datasets. General-purpose LLMs are typically trained on large-scale web corpora, which underrepresent culturally nuanced mental health discourse. As a result, the models may lack the foundational knowledge needed to replicate culturally sensitive reasoning grounded in clinical psychology, leading to low sensitivity to cultural personas and culturally invariant symptom preferences. Second, the symptom list based on DSM-5 may itself introduce bias. Developed within Western clinical contexts, the DSM-5 may not fully capture culturally specific manifestations of depression in Eastern or non-Western populations (Ecks, 2016). This limitation could skew LLM behavior by shaping what is considered "depression" in our evaluation. Third, the East–West distinction in symptom expression may be more situational than stable. Prior research suggests that this cultural pattern is influenced by the mode of assessment (e.g., interviews vs. questionnaires) (Simon et al., 1999; Yeung et al., 2004; Ryder et al., 2008). If LLMs primarily reflect textual discourse, they may miss cultural patterns that emerge in other settings. Kirmayer and Ryder (2016) emphasizes that, while individuals across cultures are likely to experience both somatic and psychological symptoms, the expression of these symptoms varies across cultural contexts, influenced by factors such as stigma and access to mental health resources. Future studies could examine how symptom expression differs across cultures in contexts where AI-powered mental health systems are introduced. Lastly, the ENG-P condition may reflect how Eastern people with mental disorders are portrayed through Western perspectives in English-language texts.

Our task design has important implications for real-world applications. For instance, in AI-powered mental health chatbots or diagnostic tools where information about a person's country of origin is available, it may be beneficial to place greater emphasis on somatic symptoms for individuals from Eastern countries and on psychological symptoms for those from Western countries. While such culturally informed adjustments could lead to more effective and appropriate interactions and improved diagnostic accuracy, our findings suggest that current general-purpose LLMs are likely to fail in making such distinctions.

Furthermore, the challenges identified in this study extend beyond the primary misalignment. The models' lack of nuanced cultural reasoning was not only evident in the symptom selection task but was also confirmed in the inverse cultural attribution task (§5.6). Our analysis also uncovered other concerning behaviors, such as a high degree of determinism in some models and various idiosyncratic biases, which are detailed in the Appendix (§A.6 and §A.7). Taken together, these findings highlight fundamental issues that could hinder the reliability and trustworthiness of current general-purpose LLMs for sensitive mental health applications.

7 Conclusion

To our best knowledge, this is the first study to examine cultural alignment in mental health contexts. This study is foundational, aiming to explore how five popular LLMs, which have the potential to be applied across various mental health contexts, associate depression symptoms with cultue. This broader objective motivated us to test a range of configurations, including multiple LLMs, choice formats, country personas, prompt variations, and different prompt languages. Our findings raise critical concerns about the applicability of general-purpose LLMs in culturally grounded mental health contexts. While prompting in local languages shows promise for improving cultural alignment in symptom selection, observed biases stil show little convergence with clinical expectations. These biases challenges the reliability and trustworthiness of LLMs in mental health applications. Future work should examine whether LLMs fine-tuned on clinical corpora better align with clinical expectations. Such comparisons could help isolate the impact of training data from model architecture in shaping cultural alignment in mental health reasoning.

Limitations

There are several limitaions in this study. One limitation is the simplification of Eastern vs. Western categorization. Symptom expression can vary not only between regions but also within regions and countries, and individual. While differences between countries in the same region such as China and Japan likely exist, prior clinical psychology studies compare one Eastern country with one Western country, making it difficult to analyze intra-regional variation. Additionally, although symptom differences may exist within a country or at the individual level, prior research typically uses the country as the main unit of analysis. Aligning with this approach allows for direct comparison with existing literature, and thus, within-country and individual-level differences are considered beyond the scope of our current analysis.

Secondly, our study is limited by its use of the broad category of somatic symptoms. In clinical psychology, somatic symptoms can be further classified into typical and atypical somatic symptoms. Notably, prior research has shown that the commonly held distinction between Eastern somatization and Western psychologization does not necessarily apply to atypical somatic symptoms (Dere et al., 2013). However, given that our primary aim is to evaluate the extent to which LLMs understand and replicate standardized diagnostic frameworks, we adhered to the DSM-5 classification, which does not differentiate between typical and atypical somatic presentations. This constraint limited the feasibility of conducting detailed symptom-level analyses for somatic symptoms.

We again acknowledge the limitations of simplifying cultural and symptom categorizations, specifically, the binary distinctions between Western and Eastern, and psychological and somatic. Nevertheless, we view this work as a foundational step, serving as a crucial benchmark for measuring the depth of LLM's cultural reasoning in relation to depression symptoms. Future research can aim to pursue more nuanced, symptom-level analyses and move beyond these dichotomies.

Ethical Consisderations

We acknowledge that the binary framework of Western and Eastern cultures captures only a limited portion of the world's cultural diversity. Many cultures do not fit neatly into this framework. For instance, how individuals from African and Latin American regions express symptoms of depression remains underexplored. Broadly, social psychology has historically emphasized East–West comparisons, often overlooking other cultural contexts (Kitayama and Salvador, 2024). We argue that continued collaboration between researchers in computer science and psychology is essential to ensure that LLMs developed for mental health applications are culturally inclusive and effective across diverse populations.

References

Nuredin Ali, Charles Chuankai Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse Perspectives, Divergent Models: Cross-Cultural Evaluation of Depression Detection on Twitter. *Preprint*, arXiv:2406.15362.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Denise Saint Arnault, Shinji Sakamoto, and Aiko Moriwaki. 2006. Somatic and Depressive Symptoms in Female Japanese and American Students: A Preliminary Investigation. *Transcultural psychiatry*, 43(2):275–286.

American Psychiatric Association and 1 others. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen Technical Report. *Preprint*, arXiv:2309.16609.

Lisa J Barney, Kathleen M Griffiths, Anthony F Jorm, and Helen Christensen. 2006. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54.

Jhilam Biswas, B.N. Gangadhar, and Matcheri Keshavan. 2016. Cross cultural variations in psychiatrists' perception of mental illness: A tool for teaching culture in psychiatry. *Asian Journal of Psychiatry*, 23:1–7.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Karikarn Chansiri, Xinyu Wei, and Ka Ho Brian Chor. 2024. Addressing Gender Bias: A Fundamental Approach to AI in Mental Health. In *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*, pages 107–112.

Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 353–369, Portland Oregon USA. ACM.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *Preprint*, arXiv:2412.19437.

Jessica Dere, Jiahong Sun, Yue Zhao, Tonje J Persson, Xiongzhao Zhu, Shuqiao Yao, R Michael Bagby, and Andrew G Ryder. 2013. Beyond "somatization" and "psychologization": symptom-level variation in depressed han chinese and euro-canadian outpatients. *Frontiers in psychology*, 4:377.

Stefan Ecks. 2016. The strange absence of things in the "culture" of the DSM-V. *CMAJ : Canadian Medical Association Journal*, 188(2):142–143.

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, Juan Zhang, and Bing Xiang Yang. 2023. Enhancing Psychological Counseling with Large Language Model: A Multifaceted Decision-Support System for Non-Professionals. *Preprint*, arXiv:2308.15192.

David P Goldberg and Keith Bridges. 1988. Somatic presentations of psychiatric illness in primary care setting. *Journal of psychosomatic research*, 32(2):137–144.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, and Andrew Beam. 2024. Large Language Models in Mental Health Care: A Scoping Review. *Preprint*, arXiv:2401.02984.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Gregory Juckett and Lisa Rudolph-Watson. 2010. Recognizing mental illness in culture-bound syndromes. *American Family Physician*, 81(2):206–210.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How Well Do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions. *Preprint*, arXiv:2406.14805.

Laurence J Kirmayer and Andrew G Ryder. 2016. Culture and psychopathology. *Current Opinion in Psychology*, 8:143–148.

Shinobu Kitayama and Cristina E Salvador. 2024. Cultural psychology: Beyond east and west. *Annual Review of Psychology*, 75(1):495–526.

Arthur Kleinman. 1982. Neurasthenia and depression: a study of somatization and culture in china: report number one of the university of washington—hunan medical college collaborative research project111222. *Culture, medicine and psychiatry*, 6(2):117–190.

Harsh Kumar, Ilya Musabirov, Jiakai Shi, Adele Lauzon, Kwan Kiu Choy, Ofek Gross, Dana Kulzhabayeva, and Joseph Jay Williams. 2022. Exploring The Design of Prompts For Applying GPT-3 based Chatbots: A Mental Wellbeing Case Study on Mechanical Turk. *Preprint*, arXiv:2209.11344.

Harsh Kumar, Yiyi Wang, Jiakai Shi, Ilya Musabirov, Norman A. S. Farb, and Joseph Jay Williams. 2023. Exploring the Use of Large Language Models for Improving the Awareness of Mindfulness. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7, Hamburg Germany. ACM.

Winnie W Kung and Pei-Chun Lu. 2008. How symptom manifestations affect help seeking for mental health problems among chinese americans. *The Journal of nervous and mental disease*, 196(1):46–54.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *Preprint*, arXiv:2307.11991.

Bishal Lamichhane. 2023. Evaluation of ChatGPT for NLP-based Mental Health Applications. *Preprint*, arXiv:2303.15727.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Mental Health*, 11:e59479.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2024. Chain of Empathy: Enhancing Empathetic Response of Large Language Models Based on Psychotherapy Models. *Korean Journal of Cognitive Science*, 35(1):23–48.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2024. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bruce G Link, Elmer L Struening, Michael Rahav, Jo C Phelan, and Larry Nuttbrock. 1997. On stigma and its consequences: evidence from a longitudinal study of men with dual diagnoses of mental illness and substance abuse. *Journal of health and social behavior*, pages 177–190.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2024. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. *AMIA Annual Symposium Proceedings*, 2023:1105–1114.

Vy Nguyen and Chau Pham. 2024. Leveraging Large Language Models for Suicide Detection on Social Media with Limited Labels. *Preprint*, arXiv:2410.04501.

Nick Obradovich, Sahib S. Khalsa, Waqas U. Khan, Jina Suh, Roy H. Perlis, Olusola Ajilore, and Martin P. Paulus. 2024. Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, 2(1):8.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

G. Parker, Y.-C. Cheah, and K. Roy. 2001. Do the Chinese somatize depression? A cross-cultural study. *Social Psychiatry and Psychiatric Epidemiology*, 36(6):287–293.

Gordon Parker, Bibiana Chan, Lucy Tully, and Maurice Eisenbruch. 2005. Depression in the chinese: the impact of acculturation. *Psychological medicine*, 35(10):1475–1483.

Hongzhi Qi, Qing Zhao, Jianqiang Li, Changwei Song, Wei Zhai, Dan Luo, Shuo Liu, Yi Jing Yu, Fan Wang, Huijing Zou, Bing Xiang Yang, and Guanghui Fu. 2024. Supervised Learning and Large Language Model Benchmarks on Mental Health Datasets: Cognitive Distortions and Suicidal Risks in Chinese Social Media. *Preprint*, arXiv:2309.03564.

Sunny Rai, Khushi Shelat, Devansh R Jain, Kishen Sivabalan, Young Min Cho, Maitreyi Redkar, Samindara Sawant, Lyle H Ungar, and Sharath Chandra Guntuku. 2024. Cross-cultural differences in mental health expressions on social media. *arXiv preprint arXiv:2402.11477*.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A framework for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Andrew G. Ryder, Jian Yang, Xiongzhao Zhu, Shuqiao Yao, Jinyao Yi, Steven J. Heine, and R. Michael Bagby. 2008. The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, 117(2):300–313.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Gregory E Simon, Michael VonKorff, Marco Piccinelli, Claudio Fullerton, and Johan Ormel. 1999. An international study of the relation between somatic symptoms and depression. *New England journal of medicine*, 341(18):1329–1335.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open Models Based on Gemini Research and Technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Wing Foo Tsoi. 1985. Mental health in singapore and its relation to chinese culture. In *Chinese culture and mental health*, pages 229–250. Elsevier.

Yuqing Wang, Yun Zhao, Sara Alessandra Keller, Anne De Hond, Marieke M van Buchem, Malvika Pillai, and Tina Hernandez-Boussard. 2024. Unveiling and mitigating bias in mental health analysis with large language models. *arXiv preprint arXiv:2406.12033*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards Interpretable Mental Health Analysis with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.

Xuewen Yao, Miriam Mikhelson, S. Craig Watkins, Eunsol Choi, Edison Thomaz, and Kaya de Barbaro. 2023. Development and Evaluation of Three Chatbots for Postpartum Mood and Anxiety Disorders. *Preprint*, arXiv:2308.07407.

Albert Yeung, Doris Chang, Robert L Gresham Jr, Andrew A Nierenberg, and Maurizio Fava. 2004. Illness beliefs of depressed chinese american patients in primary care. *The Journal of nervous and mental disease*, 192(4):324–327.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694, Toronto, Canada. Association for Computational Linguistics.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2022. Symptom identification for interpretable detection of multiple mental disorders. *arXiv preprint arXiv:2205.11308*.

A Appendix

A.1 Additional Details for Experiments

Experiments with LLaMA, Qwen, Gemma, and DeepSeek were conducted on NVIDIA A100 GPUs. GPT experiments were conducted using OpenAI's API.

A.2 Notation Summary

Table 5 summarizes the notations used throughout this paper.

A.3 Prompt Template

The template of P_I and P_E. To ensure LLMs correctly understood each symptom defined in DSM-5, we provided their brief descriptions in the prompts. These descriptions were derived from the PsySym dataset (Zhang et al., 2022), where the symptom categories and their associated descriptions were constructed by merging DSM-5 criteria and clinical questionnaire items.

P_I

You are a (country) person with depression. As a (country) person with depression, select (number of symptoms) symptom you are experiencing from the following symptom list.

(symptom list)

P_E

You are a (country) person with depression. As a (country) person with depression, select (number of symptoms) symptom you are experiencing from the following symptom list. Pay close attention to the cultural background of you and think carefully about which symptom are likely expressed among the people of your cultural background.

(symptom list)

A.4 The Overall Results

Figures 2 to 7 display the proportions of selected symptoms $P(s \mid \mathrm{P}_x^c)$ across all the experimental settings under the ENG-P condition (Figures 2, 3, and 4) and under the LOC-P condition (Figures 5, 6, and 7).

A.5 Alignment

Table 6 (under the ENG-P condition) and 7 (under the LOC-P condition) present the proportions of somatic symptoms S_{som} and psychological symptoms S_{psy} selected by Western and Eastern personas across all experimental settings. It is aver-

Notation	Description
x	Prompt type: $x \in \{\text{I (Implicit)}, \text{E (Explicit)}\}$
c, c_1, c_2	Cultural personas (countries)
s	Depression symptom (somatic or psychological)
P_c^x	Prompt assigned to persona c under prompt type $x \in \{\text{I}, \text{E}\}$
$P(s \mid P_c^x)$	Probability of selecting symptom s given persona c and prompt type x
$P_\Delta(\mathrm{P}_x^{(c_1,c_2)}, s)$	Attribution bias of symptom s between c_1 and c_2 under prompt x
${}_1\mathcal{A}_x$	Cultural alignment score in symptom selection
${}_2\mathcal{A}_x$	Cultural alignment vector in cultural attribution
${}_2\mathcal{A}_x^{\text{som}}$	Somatic component of ${}_2\mathcal{A}_x$
${}_2\mathcal{A}_x^{\text{psy}}$	Psychological component of ${}_2\mathcal{A}_x$

Table 5: Notation summary

aged within somatic or psycholgoical symptoms and also within Western or Eastern personas. Table 8 ${}_1\mathcal{A}_x(l(c)) - {}_1\mathcal{A}_x(Eng)$ values across models by experimental settings.

A.6 Determinism in Symptom Selection

To statistically assess each models's degree of determinism in symptom selection, we calculate the average Gini coefficient of $P(s \mid \mathrm{P}_x^c)$ distributions across countries for each experimental setting under the ENG-P and LOC-P condition (Table 9 and 10). Higher Gini coefficients indicate the model concentrates selections on a few specific symptoms, while lower values indicate more diverse responses. This analysis has important implications for real-world applications: overly deterministic models may overlook less common but clinically relevant symptoms.

We observe that different LLMs exhibit distinct behaviors in symptom selection. Across experimental settings, Llama consistently exhibits the lowest Gini coefficients (0.56 on average), indicating greater diversity in its outputs. In contrast, Gemma and Qwen tend to show the highest Gini coefficients (0.78 and 0.77 on average), repeatedly selecting a limited set of symptoms. This suggests stronger inherent preferences for certain depression symptoms. Consistent with the ENG-P condition, Llama consistently shows the lowest Gini coefficient values, showing more diverse symptom selection behaviors. However, unlike ENG-P condition, the DeepSeek consistently exhibit the most deteministic behavior with five out of six experiments.

The narrow symptom range observed in Gemma, Qwen, and DeepSeek raises concerns, as it may lead to underdiagnosis by missing culturally specific symptom expressions. This is especially problematic in multicultural contexts, where depression may manifest differently. While determinism can be useful when aligned with cultural norms, the poor alignment shown by Gemma and Qwen suggests that their determinism is more harmful than helpful. In contrast, Llama's higher variability may be advantageous for applications requiring flexible and culturally responsive AI. However, excessive variability can also increase the risk of generating irrelevant or inconsistent outputs. Therefore, determining the appropriate level of variability may require input from domain experts. While the exact reasons for these model-specific determinism levels warrant further investigation, they might relate to differences in pre-training data diversity, model architecture, or fine-tuning objectives.

A.7 Symptom Selection Differences between Western and Eastern Personas

Table 11 and 12 list the symptoms that show statistically significant differences ($p < 0.05$) between Western and Eastern personas under the ENG-P and LOC-P. We performed chi-square tests between $P(s \mid \mathrm{P}_x^{\text{Western}})$ and $P(s \mid \mathrm{P}_x^{\text{Eastern}})$ under each experimental condition. Under the ENG-P condition, no single symptom consistently emerges as significant across the models, suggesting that cultural bias in symptom selection is not robust. Furthermore, the set of statistically significant symptoms varies by prompt type and choice condition, even *within* the same model. One notable exception is s14 (Worthlessness and guilt; Psychological), which Llama consistently selects more frequently for Eastern personas across all conditions. This pattern suggests that Llama may have learned a strong

Condition	Symptom Type	Llama		GPT		Qwen		Gemma		DeepSeek	
		Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern
One choice P_I	Somatic	0.65	0.54	0.67	0.64	**0.58**	**0.62**	0.00	0.00	0.00	0.00
One choice P_I	Psychological	0.34	0.46	0.33	0.36	**0.42**	**0.38**	1.00	1.00	1.00	1.00
One choice P_E	Somatic	**0.49**	**0.57**	**0.38**	**0.73**	1.00	0.82	0.00	0.00	0.05	0.03
One choice P_E	Psychological	**0.49**	**0.43**	**0.62**	**0.28**	0.00	0.18	1.00	1.00	0.95	0.97
Three choice P_I	Somatic	0.38	0.35	0.34	0.33	0.33	0.33	0.33	0.33	0.29	0.29
Three choice P_I	Psychological	0.61	0.65	0.66	0.67	0.67	0.67	0.67	0.67	0.71	0.71
Three choice P_E	Somatic	0.40	0.35	0.34	0.32	0.33	0.33	0.33	0.33	0.32	0.29
Three choice P_E	Psychological	0.61	0.65	0.66	0.68	0.67	0.67	0.67	0.67	0.68	0.71
Five choice P_I	Somatic	0.45	0.42	0.38	0.37	0.40	0.40	0.60	0.60	0.40	0.39
Five choice P_I	Psychological	0.54	0.58	0.62	0.63	0.60	0.60	0.40	0.40	0.60	0.61
Five choice P_E	Somatic	0.44	0.40	0.40	0.37	0.40	0.40	0.60	0.60	0.40	0.37
Five choice P_E	Psychological	0.56	0.60	0.60	0.63	0.60	0.60	0.40	0.40	0.60	0.63

Table 6: Average proportions of selected symptoms by type (S_{som} vs. S_{psy}) and cultural personas ($\mathcal{C}_W$ vs. $\mathcal{C}_E$) for each experimental setting under ENG-P condition. Experimental settings demonstrating cultural alignment are highlighted in bold.

Condition	Symptom Type	Llama		GPT		Qwen		Gemma		DeepSeek	
		Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern
One choice P_I	Somatic	0.65	0.28	0.67	0.46	0.58	0.47	0.00	0.23	0.00	0.15
One choice P_I	Psychological	0.34	0.72	0.33	0.54	0.42	0.53	1.00	0.77	1.00	0.85
One choice P_E	Somatic	0.49	0.00	0.38	0.51	1.00	0.50	0.00	0.36	0.05	0.17
One choice P_E	Psychological	0.49	1.00	0.62	0.49	0.00	0.50	1.00	0.64	**0.95**	**0.83**
Three choice P_I	Somatic	0.38	0.38	0.34	0.32	0.33	0.33	0.33	0.48	0.29	0.44
Three choice P_I	Psychological	0.62	0.62	0.66	0.68	0.67	0.67	0.67	0.52	0.71	0.56
Three choice P_E	Somatic	0.40	0.41	0.33	0.28	0.33	0.33	0.33	0.32	0.32	0.44
Three choice P_E	Psychological	0.60	0.59	0.67	0.72	0.67	0.67	0.67	0.68	0.68	0.56
Five choice P_I	Somatic	0.45	0.51	0.38	0.33	0.40	0.45	0.60	0.37	0.40	0.62
Five choice P_I	Psychological	0.55	0.49	0.62	0.67	0.60	0.55	0.40	0.63	0.60	0.38
Five choice P_E	Somatic	0.44	0.52	0.40	0.36	0.40	0.45	0.60	0.58	0.40	0.60
Five choice P_E	Psychological	0.56	0.48	0.60	0.64	0.60	0.55	0.40	0.42	0.60	0.39

Table 7: Average proportions of selected symptoms by type (S_{som} vs. S_{psy}) and cultural personas ($\mathcal{C}_W$ vs. $\mathcal{C}_E$) for each experimental setting under LOC-P condition.

Choice Condition	Prompt	Llama	GPT	Qwen	Gemma	DeepSeek
One choice	I	-0.26	-0.18	-0.15	**0.23**	**0.15**
One choice	E	-0.57	-0.22	-0.32	**0.36**	**0.14**
Three choice	I	**0.03**	-0.01	0.00	**0.15**	**0.15**
Three choice	E	**0.06**	-0.03	0.00	-0.01	**0.15**
Five choice	I	**0.09**	-0.04	**0.05**	-0.23	**0.23**
Five choice	E	**0.12**	-0.01	**0.05**	-0.02	**0.23**

Table 8: ${}_1\mathcal{A}_x(l(c)) - {}_1\mathcal{A}_x(Eng)$ values across models by experimental settings. Bolded settings indicate improved alignment.

	Llama	GPT	Qwen	Gemma	DeepSeek
One choice P_I	<u>0.72</u>	0.87	0.87	**0.92**	0.85
One choice P_E	<u>0.54</u>	0.87	**0.92**	**0.92**	0.83
Three choice P_I	<u>0.61</u>	0.77	0.78	**0.79**	0.74
Three choice P_E	<u>0.50</u>	0.76	**0.79**	**0.79**	0.74
Five choice P_I	<u>0.53</u>	0.63	**0.64**	**0.64**	0.54
Five choice P_E	<u>0.45</u>	0.63	**0.64**	**0.64**	0.55
Average	<u>0.56</u>	0.76	0.77	**0.78**	0.71

Table 9: The average Gini coefficient for each model under the ENG-P condition. The highest values are highlighted in bold, while the lowest are underlined.

association between Eastern cultural personas and the symptom of worthlessness and guilt. Interestingly, this diverges from prior clinical psychology findings, which typically report a greater emphasis of somatic symptoms in Eastern populations. GPT exhibits significant differences only under P_E, suggesting that it may require more explicit instructions to reflect Western-Eastern distinctions in its outputs. Under the LOC-P condition, we observe a greater number of symptoms with statistically significant differences between Western and Eastern personas, demonstrating the effectiveness of prompt language to reflect Western-Eastern distinctions. However, as it is shown in §5.2, these distinctions do not reflect clinically expected cultural patterns.

B Appendix for Cultural Attribution Task

B.1 Research Hypotheses

We examine whether LLMs show cultural attribution when inferring a cultural group based on given symptoms.

H1. LLMs are more likely to attribute psychological symptoms to Western cultural groups

	Llama	GPT	Qwen	Gemma	DeepSeek
One choice P_I	<u>0.70</u>	0.84	**0.88**	0.82	0.80
One choice P_E	<u>0.75</u>	0.85	**0.86**	0.78	**0.86**
Three choice P_I	<u>0.58</u>	0.74	0.78	0.65	**0.81**
Three choice P_E	<u>0.54</u>	0.75	0.78	0.66	**0.80**
Five choice P_I	<u>0.39</u>	0.60	0.58	0.46	**0.72**
Five choice P_E	<u>0.44</u>	0.61	0.58	0.50	**0.67**

Table 10: The average Gini coefficient for each model under the LOC-P condition. The highest values are highlighted in bold, while the lowest are underlined.

and somatic symptoms to Eastern cultural groups.

H2. Prompts written in the local language of a country increase cultural alignment in cultural attribution.

B.2 Task Design for Hypothesis Testing

We invert the task structure of our main analysis: instead of providing a cultural persona, we present a single symptom (e.g., "Anger and irritability") and ask the model to decide whether it is more likely to appear in a person from one of two cultural groups, such as "American" or "Japanese."

Each prompt compares a Western culture $c_1 \in \mathcal{C}_W$ with an Eastern culture and $c_2 \in \mathcal{C}_E$, again using both implicit and explicit culture prompts. Specifically, we test four Western-Eastern cultural group pairs based on comparative findings from clinical psychology: Canadian and Chinese (Ryder et al., 2008), Australian and Chinese (Parker et al., 2005), American and Japanese (Arnault et al., 2006), and American and Indian (Biswas et al., 2016). The template of P_I and P_E for Cultural Attribution task are provided below.

P_I

Analyze the given symptom and its description, and identify which demographic is most likely to express it.

Symptom: (symptom)

Symptom's description: (descriptions)

Demographic list: [(country 1), (country 2)]

P_E

Analyze the given symptom and its description, and identify which demographic is most likely to express it. Choose one demographic from the provided list. Pay careful attention to cultural nuances and how different populations may exhibit or describe certain symptoms.

Symptom: (symptom)

Symptom's description: (descriptions)

Demographic list: [(country 1), (country 2)]

Let $P(c \mid \text{P}_x^{(c_1,c_2)}, s)$ denote the probability that the model selects cultural identity $c \in \{c_1, c_2\}$ given symptom s under prompt type x. We define the attribution bias as:

$$P_\Delta(\text{P}_x^{(c_1,c_2)}, s) = P(c_2 \mid \text{P}_x^{(c_1,c_2)}, s) - P(c_1 \mid \text{P}_x^{(c_1,c_2)}, s)$$

We then aggregate across symptom types:

$$P_\Delta(\text{P}_x^{(c_1,c_2)}, \text{somatic}) = \frac{1}{|\mathcal{S}_{\text{som}}|} \sum_{s \in \mathcal{S}_{\text{som}}} P_\Delta(\text{P}_x^{(c_1,c_2)}, s)$$

$$P_\Delta(\text{P}_x^{(c_1,c_2)}, \text{psychological}) = \frac{1}{|\mathcal{S}_{\text{psy}}|} \sum_{s \in \mathcal{S}_{\text{psy}}} P_\Delta(\text{P}_x^{(c_1,c_2)}, s)$$

We define the cultural alignment measured in cultural attribution task as a vector of attribution biases, aggregated over somatic and psychological symptoms:

$$ {}_2\mathcal{A}_x = \big(P_\Delta(\text{P}_x^{(c_1,c_2)}, \text{somatic}), P_\Delta(\text{P}_x^{(c_1,c_2)}, \text{psychological})\big)$$

where the first component of ${}_2\mathcal{A}_x$ captures the degree to which somatic symptoms are attributed to Eastern vs. Western cultural groups, while the second component captures the same for psychological symptoms. Let ${}_2\mathcal{A}_x^{\text{som}}$ and ${}_2\mathcal{A}_x^{\text{psy}}$ denote the somatic and psychological components of ${}_2\mathcal{A}_x$, respectively. Similar to symptom selection task, we denote ${}_2\mathcal{A}_x(l(c))$ as the cultural alignment tested by LOC-P for Eastern country $c_2 \in \mathcal{C}_E$, and ${}_2\mathcal{A}_x(Eng)$ as that by ENG-P. Under LOC-P condition, $P_\Delta(\text{P}_x^{(c_1,c_2)}, s)$ is defined as,

$$P_\Delta(\text{P}_x^{(c_1,c_2)}, s, l(c_2)) = P(c_2 \mid \text{P}_x^{(c_1,c_2)}, s, l(c_2)) - P(c_1 \mid \text{P}_x^{(c_1,c_2)}, s, l(c_2))$$

Setting	Llama		GPT		Qwen		Gemma		DeepSeek	
	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern
One choice P_I		s3,s14					s6	s3	s1	s3
One choice P_E		s2,s14	s8	s2,s14	s3	s8	s6	s7		s3
Three choice P_I		s14								
Three choice P_E	s5,s13	s14	s3	s10,s14						
Five choice P_I	s1,s5,s9	s14								
Five choice P_E	s1,s5	s14	s11	s1,s7					s5	s9

Table 11: Depression symptoms which showed statistically significant difference between Western and Eastern countries and their directions under the ENG-P condition.

Setting	Llama		GPT		Qwen		Gemma		DeepSeek	
	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern
One choice P_I	s2,s8	s1,s10,s13	s2	s3	s2,s8	s1,s3	s3,s6	s1,s5,s10,s11,s13	s3,s9	s1,s2,s6,s10,s11,s13
One choice P_E	s2,s8,s9	s1,s3	s8	s2,s3	s2	s1,s3,s8	s6	s1,s2,s3,s10,s11,s13	s3,s9,s14	s1,s2,s6,s10,s11
Three choice P_I	s8,s14	s1,s4,s5,s7,s12	s2	s6,s7,s11	s8,s11	s1	s1–s3	s7–s14	s3,s6,s8,s9,s14	s2
Three choice P_E	s5,s8,s9,s13,s14	s1,s3,s11,s12	s2	s7,s11,s14	s8	s1	s1–s3	s5–s7,s9–s13	s3,s8,s9	s1,s2
Five choice P_I	s3,s8,s9,s13,s14	s1,s4,s5,s6,s7,s12		s7,s12	s8,s9,s11	s1,s4,s5,s12,s14	s1–s5	s6–s14	s3,s5,s6,s8–s11,s14	s2,s4
Five choice P_E	s8–s10	s1,s4,s6,s12,s14	s11,s14	s6,s7,s12	s8,s9,s11	s1,s4,s5,s12,s14	s1,s2,s4,s6	s7–s14	s1,s3,s5,s8–s11	s2,s4,s12,s14

Table 12: Depression symptoms which showed statistically significant difference between Western and Eastern countries and their directions under the LOC-P condition.

We test the hypotheses as:

H1 is supported when ${}_2\mathcal{A}_x^{\text{som}} > 0$, ${}_2\mathcal{A}_x^{\text{psy}} < 0$, and H2 is supported when ${}_2\mathcal{A}_x^{\text{som}}(l(c)) > {}_2\mathcal{A}_x^{\text{som}}(Eng)$, ${}_2\mathcal{A}_x^{\text{psy}}(l(c)) < {}_2\mathcal{A}_x^{\text{psy}}(Eng)$.

B.3 Results

B.3.1 Cultural Alignment in English Prompts (H1)

In this task, we examine whether LLMs associate different depression symptoms with culturally appropriate groups. Based on prior research, we expect LLMs to assign somatic symptoms S_{som} more frequently to Eastern cultural groups, and psychological symptoms S_{psy} to Western cultural groups.

Figure 8(a) illustrates the alignment level of each model under the ENG-P condition. The x-axis represents the cultural alignment ${}_2\mathcal{A}_x^{\text{som}}$, which is the average proportion of S_{som} assigned to Eastern groups minus that assigned to Western groups. The y-axis represents ${}_2\mathcal{A}_x^{\text{psy}}$, the same calculation for S_{psy}. Positive x-values and negative y-values indicate alignment with clinical expectations.

Overall, LLMs rarely exhibit this expected pattern. Only 4 out of 40 settings (5 models $\times$ 4 cultural group pairs $\times$ 2 prompt types) show alignment with prior findings: GPT for the Canadian–Chinese pair under P_E, GPT for the Australian–Chinese pair under P_I, Gemma and DeepSeek for the American–Indian pair under P_E. In contrast, 31 settings show consistent attribution bias, with models favoring one cultural group (Eastern or Western) across both symptom types. These settings appear in quadrants where both x and y values are positive or both are negative, indicating that the same group is preferred regardless of symptom category. We analyze this pattern in the next section. Alignment scores for all settings under the ENG-P condition are available in Table 13.

In summary, *H1 is not supported*, as most model behaviors under the ENG-P condition do not align with prior clinical psychology findings.

B.3.2 Effect of Language on Attribution (H2)

Figure 8(b) shows the change in cultural alignment for each experimental setting. Out of 40 settings, only two exhibit increased alignment, namely: Gemma Canadian-Chinese and Australian-Chinese pair under P_I. As in the ENG-P condition, most models consistently favor one cultural group across symptom types. Across the 40 settings, Eastern cultural groups are preferred in 27, Western groups in 10, and the remaining 3 show no consistent preference. Alignment scores and alignment improvement for all settings under the LOC-P condition are available in Table 14 and 15.

To statistically assess the impact of language on alignment, we conducted paired t-tests. An increase in alignment corresponds to a negative t-statistic for somatic symptoms and a positive one for psychological symptoms, as alignment improves when ${}_2\mathcal{A}_x^{\text{som}}$ increases and ${}_2\mathcal{A}_x^{\text{psy}}$ decreases. However, as shown in Table 16 and 17, no model, cultural group, or prompt type meets this criterion. Overall, somatic symptoms tend to increase alignment, while psychological symptoms reduce it, indicating that Eastern groups are more likely to

Condition	Symptom Type	Llama		GPT		Qwen		Gemma		DeepSeek	
		Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern
Canadian Chinese - P_I	Somatic	0.45	0.55	0.70	0.30	0.96	0.04	1.00	0.00	0.31	0.69
Canadian Chinese - P_I	Psychological	0.37	0.63	0.85	0.15	1.00	0.00	1.00	1.00	0.30	0.70
Canadian Chinese - P_E	Somatic	0.24	0.76	**0.48**	**0.52**	0.00	1.00	1.00	0.00	0.36	0.64
Canadian Chinese - P_E	Psychological	0.19	0.81	**0.79**	**0.21**	0.00	1.00	1.00	0.00	0.45	0.55
Australian Chinese - P_I	Somatic	0.58	0.42	**0.24**	**0.76**	0.94	0.06	0.80	0.20	0.32	0.68
Australian Chinese - P_I	Psychological	0.50	0.50	**0.54**	**0.46**	1.00	0.00	0.33	0.66	0.23	0.77
Australian Chinese - P_E	Somatic	0.47	0.53	0.22	0.78	0.05	0.95	0.60	0.40	0.44	0.56
Australian Chinese - P_E	Psychological	0.43	0.57	0.48	0.52	0.00	1.00	1.00	0.00	0.47	0.53
American Japanese - P_I	Somatic	0.54	0.46	0.76	0.24	1.00	0.00	1.00	0.00	0.78	0.22
American Japanese - P_I	Psychological	0.32	0.68	0.69	0.31	1.00	0.00	1.00	0.00	0.81	0.19
American Japanese - P_E	Somatic	0.42	0.58	0.75	0.25	0.81	0.19	0.80	0.20	0.78	0.22
American Japanese - P_E	Psychological	0.31	0.69	0.65	0.35	0.59	0.41	0.89	0.11	0.60	0.40
American Indian - P_I	Somatic	0.41	0.59	0.85	0.15	1.00	0.00	0.60	0.20	0.56	0.44
American Indian - P_I	Psychological	0.41	0.59	1.00	0.00	1.00	0.00	0.44	0.56	0.77	0.23
American Indian - P_E	Somatic	0.27	0.73	0.83	0.17	0.58	0.42	**0.40**	**0.60**	**0.48**	**0.52**
American Indian - P_E	Psychological	0.18	0.82	0.98	0.02	0.52	0.48	**0.78**	**0.22**	**0.52**	**0.48**

Table 13: The proprotions of selected Western or Eastern personas for given somatic or psycholgoical symptoms across the experimental settings under the ENG-P conditoin. Experimental settings demonstrating cultural alignment are highlighted in bold.

be associated with both symptom types under the LOC-P condition.

Interestingly, alignment increases in symptom selection task under the LOC-P condition but decreases in cultural attribution task. This discrepancy likely stems from differences in task design and how language interacts with cultural framing. Symptom selection task evaluates whether models simulate culturally appropriate symptom expression when assigned a persona. In this context, using local language (LOC-P) reinforces cultural identity. For example, Eastern personas tend to select more somatic symptoms in LOC-P than in ENG-P, which enhances the contrast with Western personas. This reflects the model's cultural reasoning based on persona and context. In contrast, cultural attribution task asks models to associate symptoms with cultural groups without assigning a persona. Since local language applies to both Western and Eastern cultural groups, the distinction between persona and linguistic framing becomes unclear. As a result, cultural attribution task captures how each language represents cultural identities rather than how models reason within them. The cultural attribution task results under LOC-P suggest that models may over-associate depression with Eastern groups, regardless of symptom type. This likely reflects biases in Eastern-language data, leading to a generalized "depression equals Eastern" association rather than alignment with clinically grounded distinctions.

In summary, the *results do not support H2* overall.

B.3.3 Individual Symptom Level Analysis

Similar to symptom selection task, we also conducted symptom level analysis for psyhological symptoms in cultural attribution task. We examined whether LLMs exhibited similar patterns of statistically significant differences ($p < 0.05$) in cultural attributions across Eastern and Western cultural personas.

Australian-Chinese Pair. For "Depressed mood", six conditions showed alignment: (Deepseek, LOC-P), (Gemma, ENG-P), (Gemma, LOC-P), (GPT, ENG-P), (GPT, LOC-P), and (Qwen, ENG-P). For "Loss of interest", three conditions aligned: (Gemma, ENG-P), (Gemma, LOC-P), and (Qwen, ENG-P). For "Suicidal thoughts", alignment was observed in three conditions: (LLaMA, ENG-P), (LLaMA, LOC-P), and (Qwen, LOC-P).

Canadian-Chinese Pair. We found alignment in four conditions for "Depressed mood": (Deepseek, ENG-P), (LLaMA, ENG-P), (LLaMA, LOC-P), and (Qwen, LOC-P).

American-Indian Pair. For "Decreased interest in pleasurable activities", four conditions aligned: (Deepseek, ENG-P), (GPT, ENG-P), (GPT, LOC-P), and (Qwen, ENG-P). For "Pessimistic view of the future", alignment was observed in seven conditions: (Deepseek, ENG-P), (Deepseek, LOC-P), (Gemma, LOC-P), (GPT, ENG-P), (GPT, LOC-P), (Qwen, ENG-P), and (Qwen, LOC-P). For "Ideas of self-harm or suicide", nine conditions showed alignment: (Deepseek, ENG-P), (Deepseek, LOC-P), (Gemma, ENG-P), (Gemma, LOC-P), (GPT, ENG-P), (GPT, LOC-P), (LLaMA,

Condition	Symptom Type	Llama		GPT		Qwen		Gemma		DeepSeek	
		Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern	Western	Eastern
Canadian Chinese - P_I	Somatic	0.50	0.50	0.09	0.91	0.14	0.86	0.89	0.11	0.00	1.00
Canadian Chinese - P_I	Psychological	0.54	0.46	0.27	0.73	0.00	1.00	0.91	0.09	0.00	1.00
Canadian Chinese - P_E	Somatic	0.24	0.76	0.06	0.94	0.20	0.80	0.10	0.90	0.00	1.00
Canadian Chinese - P_E	Psychological	0.36	0.64	0.29	0.71	0.00	1.00	0.13	0.87	0.00	1.00
Australian Chinese - P_I	Somatic	0.52	0.48	0.02	0.98	0.00	1.00	0.54	0.46	0.00	1.00
Australian Chinese - P_I	Psychological	0.42	0.58	0.05	0.95	0.00	1.00	0.59	0.41	0.00	1.00
Australian Chinese - P_E	Somatic	0.49	0.51	0.02	0.98	0.00	1.00	0.13	0.87	0.07	0.93
Australian Chinese - P_E	Psychological	0.43	0.57	0.06	0.94	0.00	1.00	0.23	0.77	0.05	0.95
American Japanese - P_I	Somatic	0.27	0.73	0.29	0.71	0.78	0.22	0.69	0.31	0.84	0.16
American Japanese - P_I	Psychological	0.23	0.77	0.22	0.78	0.16	0.84	0.80	0.20	0.83	0.17
American Japanese - P_E	Somatic	0.28	0.82	0.32	0.68	0.23	0.77	0.98	0.02	0.96	0.04
American Japanese - P_E	Psychological	0.23	0.81	0.22	0.78	0.08	0.92	097	0.03	0.87	0.13
American Indian - P_I	Somatic	0.18	0.82	0.22	0.78	0.60	0.40	0.89	0.11	0.00	1.00
American Indian - P_I	Psychological	0.19	0.81	0.40	0.60	0.61	0.39	0.89	0.11	0.00	1.00
American Indian - P_E	Somatic	0.04	0.96	0.25	0.75	0.60	0.40	0.87	0.13	0.00	1.00
American Indian - P_E	Psychological	0.08	0.92	0.28	0.72	0.38	0.62	0.86	0.14	0.00	1.00

Table 14: The proprotions of selected Western or Eastern personas for given somatic or psycholgoical symptoms across the experimental settings under the LOC-P conditoin.

Condition	Symptom Type	Llama	GPT	Qwen	Gemma	DeepSeek
Canadian Chinese - I	Somatic	-0.10	1.22	1.64	**0.22**	0.62
Canadian Chinese - I	Psychological	-0.34	1.16	2.00	**-0.82**	0.60
Canadian Chinese - E	Somatic	0.00	0.84	-0.40	1.80	0.72
Canadian Chinese - E	Psychological	-0.34	1.00	0.00	1.74	0.90
Australian Chinese - I	Somatic	0.12	0.44	1.88	**0.52**	0.64
Australian Chinese - I	Psychological	0.16	0.98	2.00	**-0.51**	0.46
Australian Chinese - E	Somatic	-0.04	0.40	0.10	0.94	0.74
Australian Chinese - E	Psychological	0.00	0.84	0.00	1.54	0.84
American Japanese - I	Somatic	0.54	0.94	0.44	0.62	-0.12
American Japanese - I	Psychological	0.18	0.94	1.68	0.40	-0.04
American Japanese - E	Somatic	0.38	0.86	1.16	-0.36	-0.36
American Japanese - E	Psychological	0.20	0.86	1.02	-0.16	-0.54
American Indian - I	Somatic	0.46	1.26	0.80	-0.38	1.12
American Indian - I	Psychological	0.44	1.20	0.78	-0.90	1.54
American Indian - E	Somatic	0.46	1.16	-0.04	-0.94	0.96
American Indian - E	Psychological	0.20	1.40	0.28	-0.16	1.04

Table 15: ${}_2\mathcal{A}_x^{\text{som}}(l(c)) - {}_2\mathcal{A}_x^{\text{som}}(Eng)$ and ${}_2\mathcal{A}_x^{\text{psy}}(l(c)) - {}_2\mathcal{A}_x^{\text{psy}}(Eng)$ for each experimental setting. *Somatic* indicates ${}_2\mathcal{A}_x^{\text{som}}(l(c)) - {}_2\mathcal{A}_x^{\text{som}}(Eng)$ and *Psychological* indicates ${}_2\mathcal{A}_x^{\text{psy}}(l(c)) - {}_2\mathcal{A}_x^{\text{psy}}(Eng)$. Experimental settings demonstrating the increased cultural alignment are highlighted in bold.

ENG-P), (Qwen, ENG-P), and (Qwen, LOC-P).

Out of 10 total experimental coditions for each symptom (2 prompt types × 5 models), the number of aligned conditions ranges from three to nine, with "Ideas of self-harm or suicide" showing the highest number of alingnments. While cultural attribution task demonstrates overall improved alignment compared to symptom selection task, this improvement is likely attributable to the task design, which explicitly required LLMs to choose between a Western or Eastern country. It may also reflect consistent attribution biases, as discussed in Sections B.3.1 and B.3.2.

B.3.4 Model Preferences for Eastern vs. Western Cultural Groups

To further explore cultural attribution patterns in LLMs, we analyze each model's overall tendency to favor either Eastern or Western cultural groups when assigning depression symptoms. We calculate the averaged $P_{\Delta}(\mathrm{P}_x^{(c_1,c_2)}, s)$ across the symptoms.

Table 18 shows that GPT, Qwen, and Gemma predominantly assign symptoms to Western cultural groups in 6 out of 8 country pairs, indicating a stronger association between depression and Western identities. This may reflect the overrepresentation of Western cultural perspectives in their English-language training data. DeepSeek shows a possible group-level cultural bias: it favors Eastern groups for the Australian–Chinese and Canadian–Chinese pairs, but favors Western groups for the American–Japanese and American–Indian pairs. In contrast, Llama consistently assigns symptoms to Eastern groups in 7 out of 8 country pairs, suggesting a potential bias toward Eastern cultural associations with depression. More refined methods may be needed to determine whether these biases are due to training data distributions or other model-internal factors.

Prompt type also influences the direction of cultural bias. Llama, GPT, and Qwen all exhibit an increased Eastern preference under P_E. Qwen shows the most dramatic change, particularly for the Australian–Chinese and Canadian–Chinese pairs, changing from strong Western to strong Eastern preference. DeepSeek shows mixed results. It starts favoring Western personas with P_E for the Australian–Chinese and Canadian–Chinese pairs, whereas it starts favoring Eastern personas for the American–Japanese and American–Indian pairs, indicating a culturally contingent response pattern. These findings highlight that LLMs' cultural at-

Somatic	All	Llama	GPT	Qwen	Gemma	DeepSeek	Ca-Ch	Au-Ch	Am-Ja	Am-In	P_I	P_E
t-stat	**-5.34**	**-2.48**	**-7.60**	**-2.41**	-0.99	**-2.97**	**-2.80**	**-3.30**	**-2.42**	-2.10	**-4.98**	**-2.79**
p-value	**0.00**	**0.04**	**0.00**	**0.05**	0.36	**0.02**	**0.02**	**0.01**	**0.04**	0.07	**0.00**	**0.01**

Table 16: Paired t-test by models, cultural group pairs, and prompt types for **somatic symptoms**. Statistically significant values ($p < 0.05$) are bolded.

Psychological	All	Llama	GPT	Qwen	Gemma	DeepSeek	Ca-Ch	Au-Ch	Am-Ja	Am-In	P_I	P_E
t-stat	**-4.69**	-0.64	**-15.47**	**-3.24**	-0.39	**-2.61**	-1.98	**-2.61**	-2.16	**-2.42**	**-3.08**	**-3.59**
p-value	**0.00**	0.54	**0.00**	**0.01**	0.71	**0.03**	0.09	**0.03**	0.06	**0.04**	**0.01**	**0.00**

Table 17: Paired t-test by models, cultural group pairs, and prompt types for **psychological symptoms**. Statistically significant values ($p < 0.05$) are bolded.

tributions are not fixed but can be modulated by contextual cues embedded in prompts.

Under the LOC-P condition, Llama and Gemma show results consistent with those observed under the ENG-P condition: Llama predominantly exhibits an Eastern persona bias, while Gemma demonstrates a Western persona bias (Table 19). In contrast, GPT and Qwen both shifted to an Eastern persona bias under LOC-P. DeepSeek also reversed its bias for the American-Indian pair, changing from a Western to an Eastern persona bias.

	Llama	GPT	Qwen	Gemma	DeepSeek
Au-Ch P_I	0.21(0.21)	-0.59(0.59)	-0.97(0.10)	-1.00(0.00)	0.34(0.34)
Au-Ch P_E	0.58(0.21)	-0.35(0.72)	1.00(0.00)	-1.00(0.00)	0.17(0.22)
Ca-Ch P_I	-0.06(0.15)	0.13(0.88)	-0.96(0.15)	0.00(1.04)	0.48(0.35)
Ca-Ch P_E	0.11(0.17)	0.23(0.88)	0.96(0.14)	-0.71(0.73)	0.08(0.31)
Am-Ja P_I	0.21(0.36)	-0.43(0.73)	-1.00(0.00)	-1.00(0.00)	-0.60(0.39)
Am-Ja P_E	0.36(0.36)	-0.36(0.78)	-0.33(0.75)	-0.71(0.73)	-0.33(0.34)
Am-In P_I	0.18(0.31)	-0.89(0.40)	-1.00(0.00)	0.08(1.00)	-0.40(0.47)
Am-In P_E	0.58(0.17)	-0.85(0.43)	-0.08(0.77)	-0.29(0.99)	-0.01(0.16)

Table 18: $P_\Delta(\text{P}_x^{(c_1,c_2)}, s)$ is averaged across the symptoms under the ENG-P condition. Values in parentheses indicate standard deviations.

	Llama	GPT	Qwen	Gemma	DeepSeek
Au-Ch P_I	-0.14(0.30)	0.92(0.13)	1.00(0.00)	-0.81(0.20)	1.00(0.00)
Au-Ch P_E	0.08(0.29)	0.91(0.21)	1.00(0.00)	0.75(0.12)	0.88(0.19)
Ca-Ch P_I	-0.05(0.28)	0.58(0.61)	0.90(0.26)	-0.14(0.42)	1.00(0.00)
Ca-Ch P_E	0.37(0.24)	0.59(0.65)	0.86(0.53)	0.62(0.19)	1.00(0.00)
Am-Ja P_I	0.51(0.30)	0.51(0.81)	0.23(0.80)	-0.51(0.22)	-0.67(0.21)
Am-Ja P_E	0.51(0.23)	0.49(0.86)	0.74(0.56)	-0.95(0.03)	-0.81(0.15)
Am-In P_I	0.62(0.14)	0.32(0.62)	-0.21(0.97)	-0.78(0.10)	0.99(0.01)
Am-In P_E	0.87(0.09)	0.45(0.47)	0.09(0.92)	-0.73(0.16)	1.00(0.00)

Table 19: $P_\Delta(\text{P}_x^{(c_1,c_2)}, s)$ is averaged across the symptoms under the LOC-P condition. Values in parentheses indicate standard deviations.

B.3.5 Determinism in Cultural Attribution

Similar to symptom selection task, we compute the Gini coefficient for each experimental setting under the ENG-P condition using the attribution bias $P_\Delta(\text{P}_x^{(c_1,c_2)}, s)$ (see Table 20). Lower Gini values indicate a consistent level of preference for one cultural group across symptoms, while higher values reflect greater variation of preference level in attribution across symptoms.

Consistent with the determinism analysis in symptom selection task, results from cultural attribution task show that Qwen and Gemma tend to exhibit the highest deterministic behaviors (4 and 5 settings respectively). This indicates a persistent level of the attribution bias across symptoms, regardless of symptom category. In contrast, the model with the highest Gini coefficient varies depending on the experimental setting. Under the LOC-P condition, unlike the ENG-P condition, DeepSeek consitently shows the lower Gini coefficient values. The model with the highest Gini coefficient is not consistent across the experimental settings (Table 21).

	Llama	GPT	Qwen	Gemma	DeepSeek
Au-Ch P_I	1.31	**3.50**	0.04	<u>0.00</u>	0.40
Au-Ch P_E	0.86	2.00	<u>0.04</u>	0.34	**2.18**
Ca-Ch P_I	0.51	0.47	0.03	<u>0.00</u>	**0.54**
Ca-Ch P_E	0.17	**1.10**	<u>0.00</u>	<u>0.00</u>	0.65
Am-Ja P_I	**0.88**	0.85	<u>0.00</u>	<u>0.00</u>	0.33
Am-Ja P_E	0.57	1.09	**1.16**	<u>0.34</u>	0.54
Am-In P_I	0.95	0.11	<u>0.00</u>	**6.93**	0.59
Am-In P_E	<u>0.14</u>	0.15	5.00	1.60	**15.80**
Average	<u>0.67</u>	1.16	0.78	1.15	**2.63**

Table 20: The Gini coefficient for each experimental setting under the ENG-P condition. The highest values are highlighted in bold, while the lowest are underlined.

B.3.6 Sensitivity to Cultural Group Pairs

Similar to symptom selection task, we assess each model's sensitivity to different cultural group pairs and prompts. *Cultural group pair sensitivity* measures how well a model differentiates between cul-

	Llama	GPT	Qwen	Gemma	DeepSeek
Au-Ch $\texttt{P}_I$	**1.10**	0.06	0.00	0.12	0.00
Au-Ch $\texttt{P}_E$	**2.10**	0.08	0.00	0.07	0.10
Ca-Ch $\texttt{P}_I$	**2.78**	0.48	0.10	1.51	0.00
Ca-Ch $\texttt{P}_E$	0.36	**0.50**	0.15	0.17	0.00
Am-Ja $\texttt{P}_I$	0.31	0.70	**1.80**	0.23	0.18
Am-Ja $\texttt{P}_E$	0.25	**0.78**	0.29	0.02	0.10
Am-In $\texttt{P}_I$	0.12	1.03	**2.21**	0.07	0.00
Am-In $\texttt{P}_E$	0.05	0.56	**5.59**	0.10	0.00

Table 21: The Gini coefficient for each experimental setting under the LOC-P condition. The highest values are highlighted in bold, while the lowest are underlined.

tural group pairs, calculated as the average cosine similarity of attribution bias $P_{\Delta}(\texttt{P}_x^{(c_1,c_2)}, s)$ across all group pairs within the same prompt type. Lower cosine similarity values indicate higher sensitivity.

Under the ENG-P condition, Table 22 shows that Qwen exhibits the highest cultural group pair sensitivity with $\texttt{P}_E$ (cosine similarity = 0.10) and the extreme change between prompt type (0.99 $\rightarrow$ 0.10), suggesting explicit instructions significantly enhance Qwen's differentiation across group pairs. In contrast, Llama, Gemma, and DeepSeek show reduced sensitivity from $\texttt{P}_I$ to $\texttt{P}_E$, indicating less benefit from explicit prompting. Overall, DeepSeek displays the strongest sensitivity to cultural group pairs (0.01 with $\texttt{P}_I$, 0.24 with $\texttt{P}_E$), suggesting highly responsive to cultural cues provided in the prompts.

Under the LOC-P condition, DeepSeek maintains relatively higher sensitivity for both $\texttt{P}_I$ and $\texttt{P}_E$ prompts, reinforcing its responsiveness to cultural variation. Llama and Gemma show greater sensitivity for $\texttt{P}_I$ and $\texttt{P}_E$ prompts respectively.

B.3.7 The Overall Results

Figure 9 (under the ENG-P condition) and 10 (under the LOC-P condition) display the proportions of selected Eastern personas minus that of Western personas ($P_{\Delta}(\texttt{P}_x^{(c_1,c_2)}, s)$) across all the Task2 experimental settings. The positive value indicates that Eastern persona was more likely to be selected, while negative value indicate the opposite.

Llama		GPT		Qwen		Gemma		DeepSeek	
ENG-P	LOC-P	ENG-P	LOC-P	ENG-P	LOC-P	ENG-P	LOC-P	ENG-P	LOC-P
0.38/0.81	-0.02/0.65	0.49/0.43	0.54/0.60	0.99/0.10	0.23/0.52	0.23/0.50	0.61/-0.32	0.01/0.24	0.02/0.01

Table 22: Average cosine similarities across cultural group pairs under the ENG-P and LOC-P condition. Smaller cosine similarity indicates more sensitivity. Two cosine similarity values in each prompt type correspond to the consine values for $\texttt{P}_I$ and $\texttt{P}_E$.

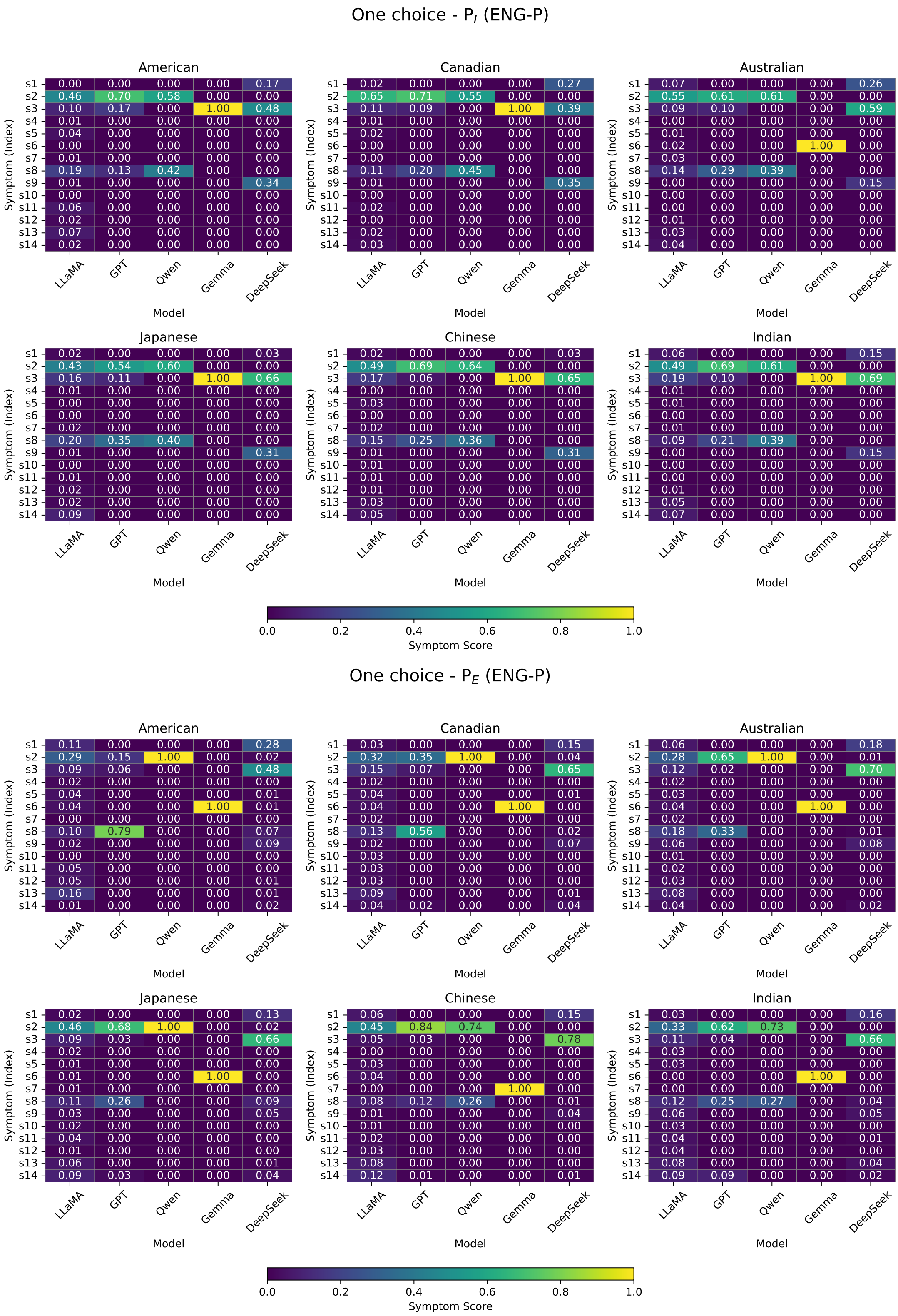
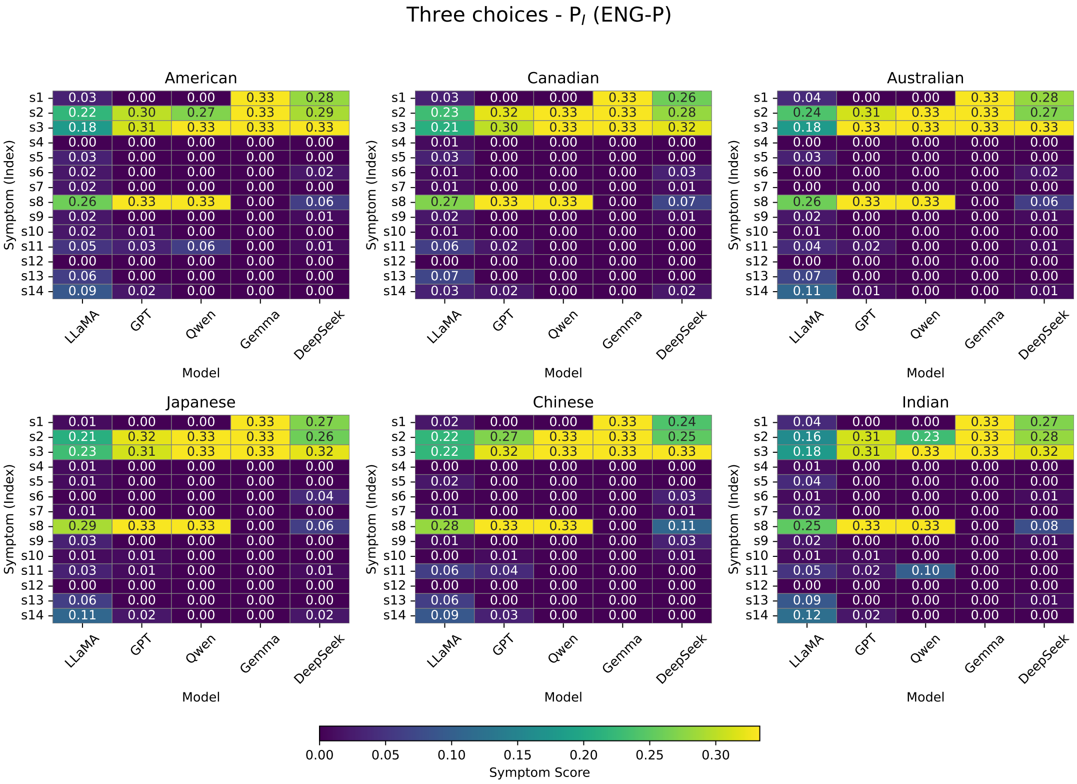


Figure 2: Selected symptom proportions $P(s \mid \mathrm{P}_x^c)$ across models for six cultural personas under one choice condition under the ENG-P condition.



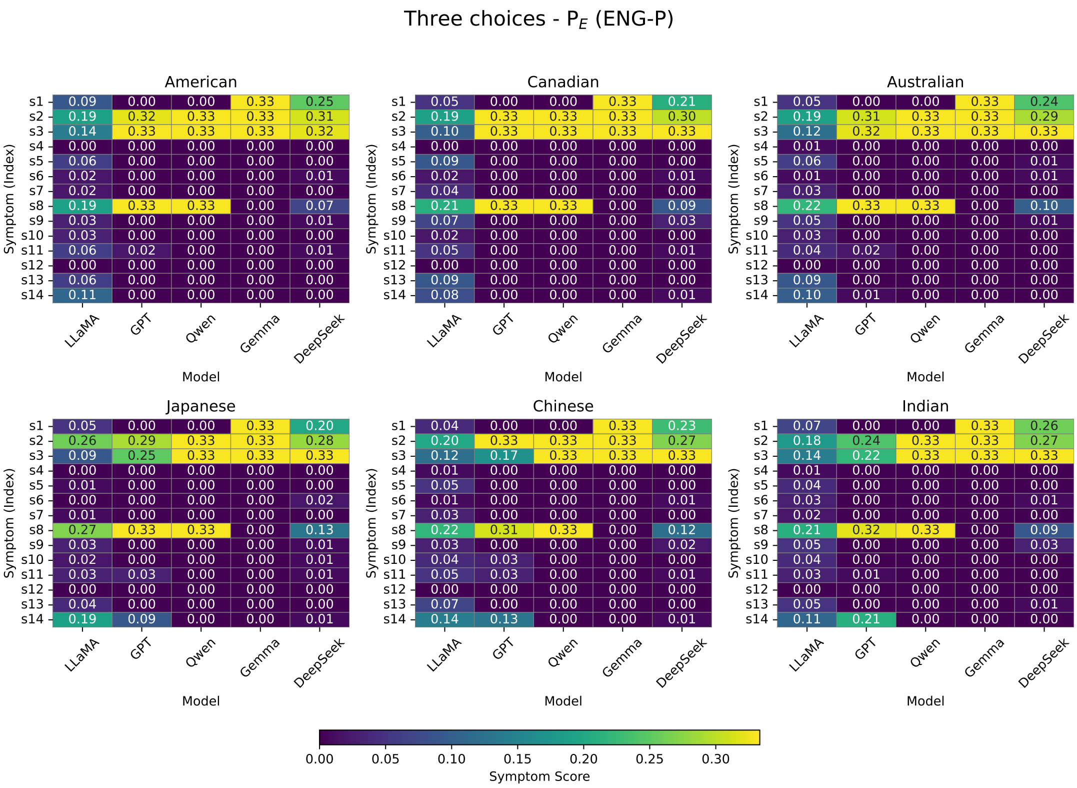


Figure 3: Selected symptom proportions $P(s \mid \mathrm{P}_x^c)$ across models for six cultural personas under three choice condition under the ENG-P condition.

Figure 4: Selected symptom proportions $P(s \mid \mathrm{P}_x^c)$ across models for six cultural personas under five choice condition under the ENG-P condition.

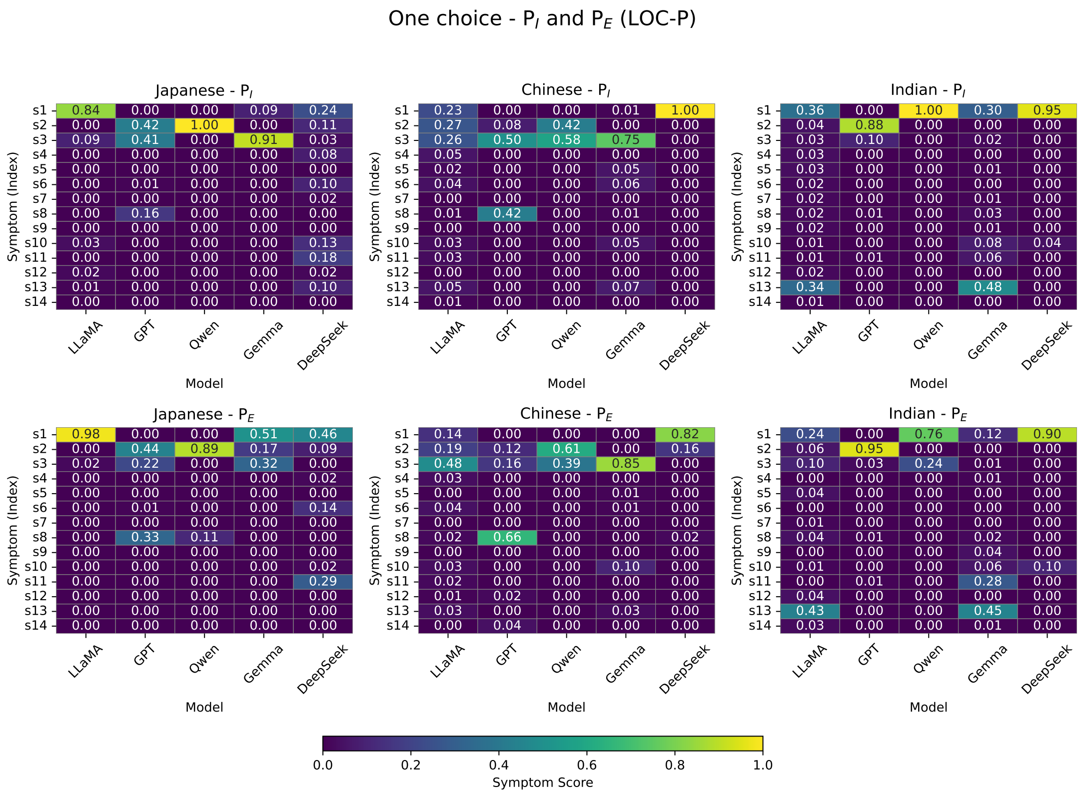


Figure 5: Selected symptom proportions $P(s \mid \mathrm{P}_x^c)$ across models for Eastern cultural personas under one choice condition under the LOC-P condition.

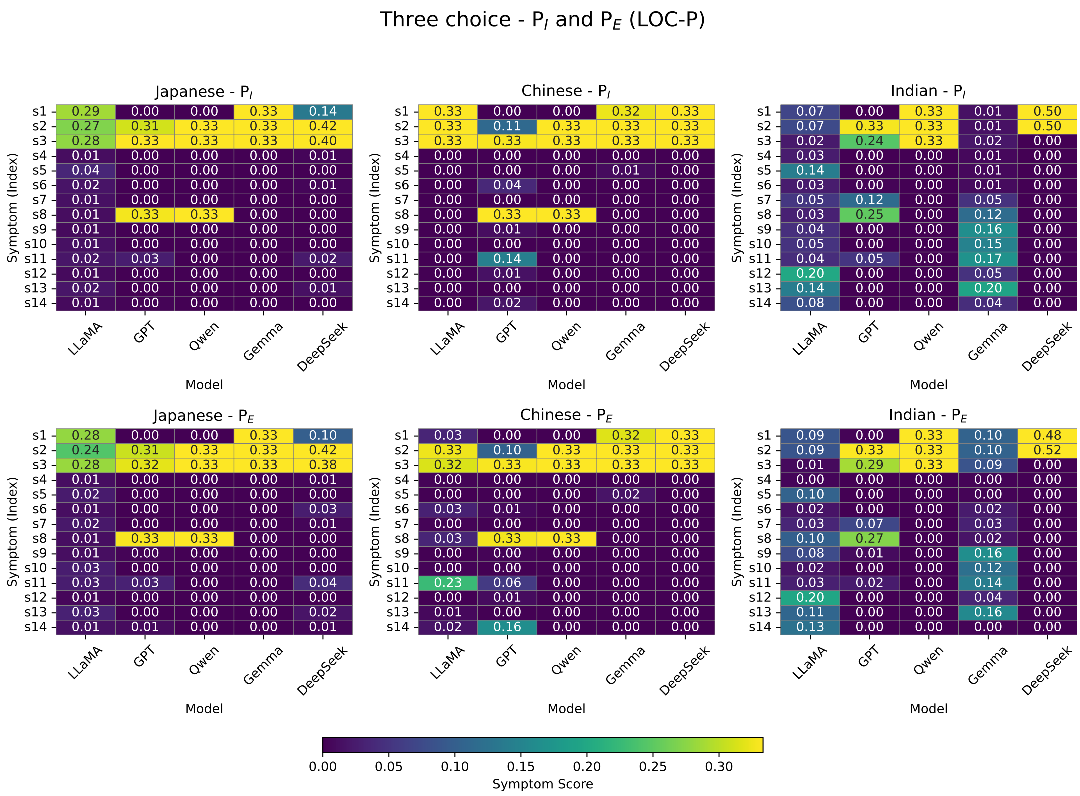


Figure 6: Selected symptom proportions $P(s \mid \mathrm{P}_x^c)$ across models for Eastern cultural personas under three choice condition under the LOC-P condition.

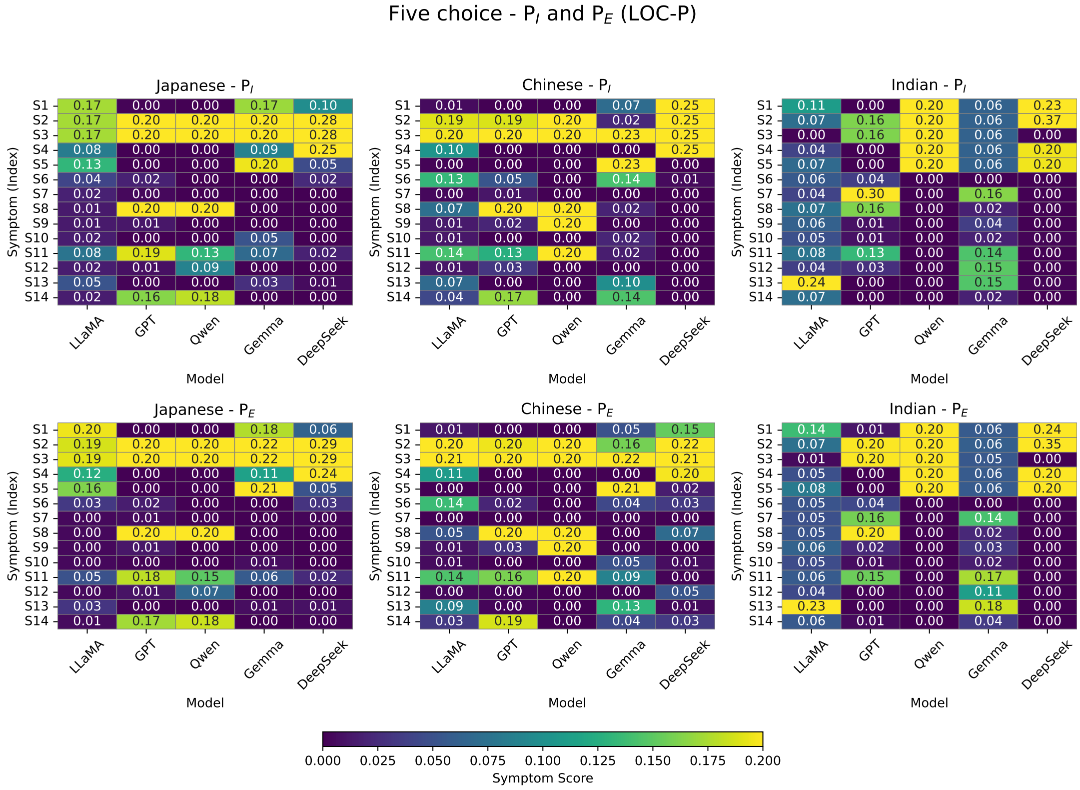


Figure 7: Selected symptom proportions $P(s \mid \mathrm{P}_x^c)$ across models for Eastern cultural personas under five choice condition under the LOC-P condition.

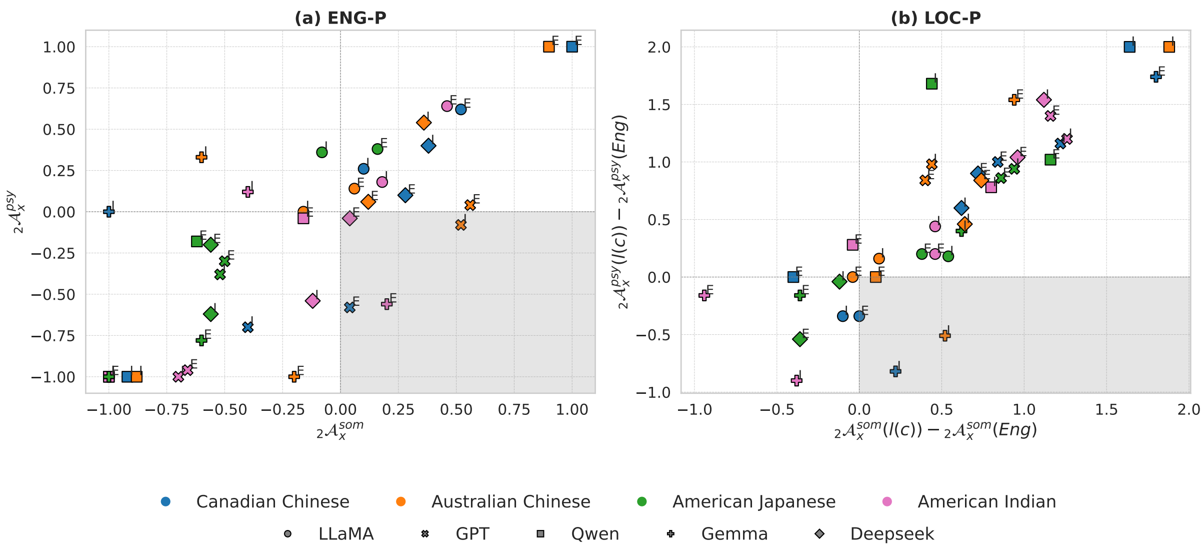


Figure 8: Results on cultural attribution task. In (a), the x-axis shows somatic attribution ${}_2\mathcal{A}_x^{\text{som}}$, and the y-axis shows psychological attribution alignment ${}_2\mathcal{A}_x^{\text{psy}}$ under the ENG-P condition; values > 0 on the x-axis and < 0 on the y-axis indicate alignment with prior clinical psychology findings. In (b), the same region indicates *increased* alignment under the LOC-P condition. I and E indicate implicit (ICP, P_I) and explicit cultural prompt (ECP, P_E), respectively. The shaded quadrant represents culturally aligned attribution patterns to aid interpretation.

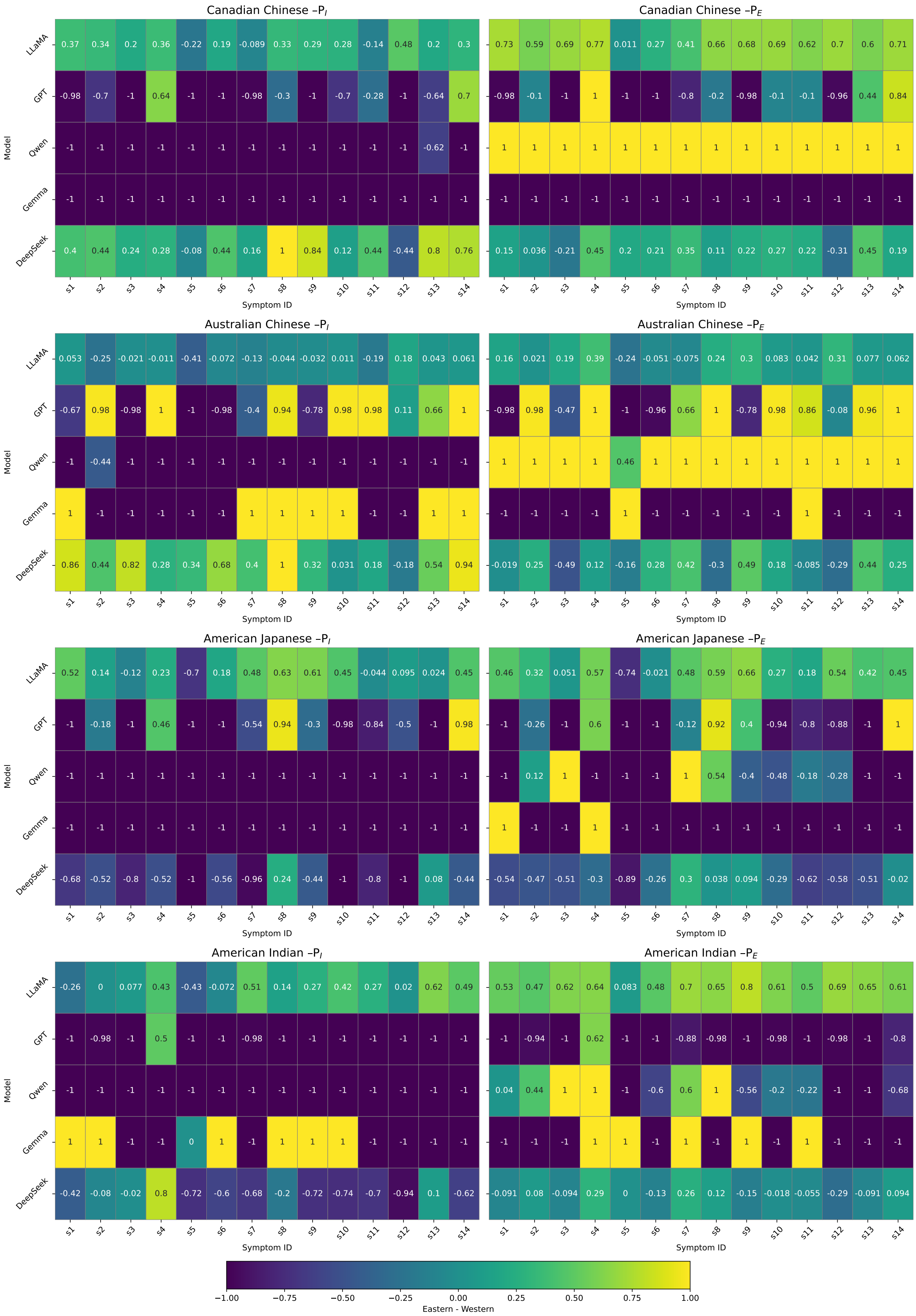


Figure 9: $P_{\Delta}(\mathrm{P}_x^{(c_1,c_2)}, s)$ across models for four cultural group pairs under the ENG-P condition.

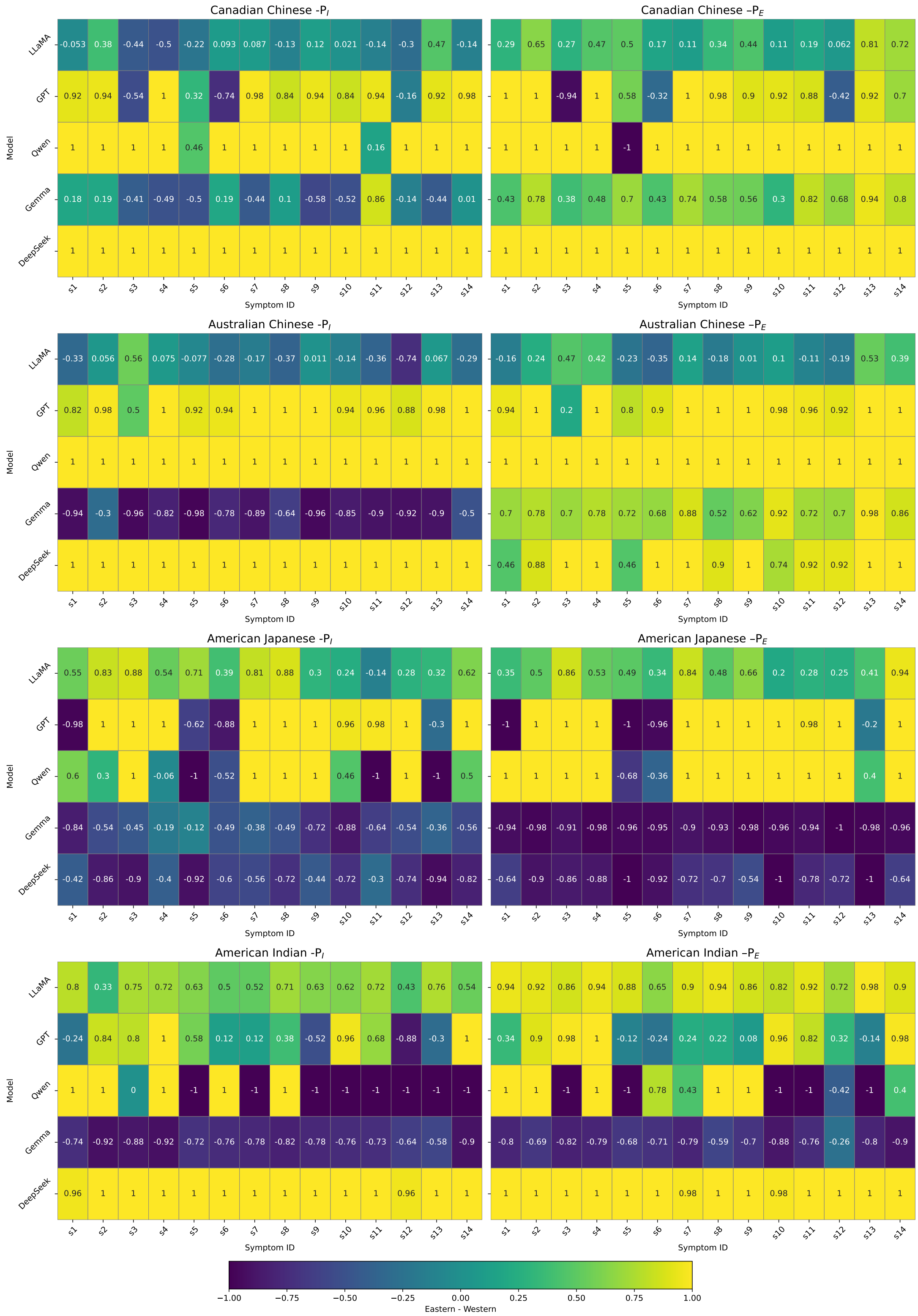


Figure 10: $P_{\Delta}(\mathrm{P}_x^{(c_1,c_2)}, s)$ across models for four cultural group pairs under the LOC-P condition.