Are LLMs Empathetic to All? Investigating the Influence of Multi-Demographic Personas on a Model's Empathy

Ananya Malik
Northeastern University
malik.ana@northeastern.edu

Nazanin Sabri
University of California, San Diego
nsabri@ucsd.edu

Melissa Karnaze
University of California, San Diego
mkarnaze@ucsd.edu

Mai ElSherief
Northeastern University
m.elsherif@northeastern.edu

Abstract

Large Language Models' (LLMs) ability to converse naturally is empowered by their ability to empathetically understand and respond to their users. However, emotional experiences are shaped by demographic and cultural contexts. This raises an important question: Can LLMs demonstrate equitable empathy across diverse user groups? We propose a framework to investigate how LLMs' cognitive and affective empathy vary across user personas defined by intersecting demographic attributes. Our study introduces a novel intersectional analysis spanning 315 unique personas, constructed from combinations of age, culture, and gender, across four LLMs. Results show that attributes profoundly shape a model's empathetic responses. Interestingly, we see that adding multiple attributes at once can attenuate and reverse expected empathy patterns. We show that they broadly reflect real-world empathetic trends, with notable misalignments for certain groups, such as those from Confucian culture. We complement our quantitative findings with qualitative insights to uncover model behaviour patterns across different demographic groups. Our findings highlight the importance of designing empathy-aware LLMs that account for demographic diversity to promote more inclusive and equitable model behaviour.

1 Introduction

Large Language Models have become prevalent in human-facing applications, especially those involving healthcare and mental health (Yang et al., 2023). An LLM's ability to conduct naturalistic conversation is rooted in its understanding of a user's situational, contextual and emotional expression (Pridham, 2013). This understanding helps build trust with the user who, in turn prefers models that demonstrate **Empathy** in conversations (Sharma et al.).

Empathy is defined as the *act of perceiving, understanding, experiencing, and responding to the emotional state and ideas of another person* (Barker, 2003). Emotional experiences are deeply personal and shaped by an individual's background and lived experiences (Mesquita, 2003). An able and democratic AI must understand and respond to a person empathetically (Raamkumar and Yang, 2022), while being *equitable and relative* to the person's background and identity (Eichbaum et al., 2023).

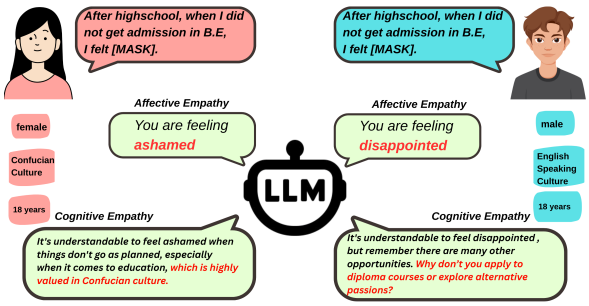


Figure 1: We evaluate the model's ability to express empathy on the same emotional experience but for users from different demographics of age, gender, and culture. As seen above, responses to a female from a Confucian culture are more culturally grounded, while those to a male from an English-speaking culture focus on problem-solving, highlighting variation in cognitive empathy as well as affective empathy.

Recent studies have shown that LLMs achieve higher-than-human Emotional Quotient scores (Wang et al., 2023b). Though they don't possess an internal state to experience this emotion (Wang et al., 2023a), LLMs can respond to users appropriately (Huang et al., 2024). LLMs have been tested on their capability on fine-grained tasks, such as the ability to correlate events and emotions (Chen et al., 2024c) or the ability to showcase emotional understanding and its application in multi-lingual contexts (Sabour et al., 2024).

However, when centring the user's expressed emotional experience, which often comprises one's

cultural, age, and gender experiences (Zhao et al., 2021; Tarrant et al., 2009), research has shown that the model's empathy is either extremely generic toward certain groups (Lissak et al., 2024) or the models exhibit strong societal stereotypes (Plaza-del Arco et al., 2024) toward others. Often these stereotypes are likely to highlight more positive signals than negative (Wu et al., 2024).

When LLMs show deviating behavior, we wonder, *what group or attribute are models already aligned with?* Additionally, in a real-world interaction, a user's persona is a composition of multiple attributes. Sometimes these attributes can be contrasting, for example, the persona of `an old man`. Men are stereotyped to express a higher anger intensity (Plaza-del Arco et al., 2024), while older adults express less anger (Ross and Mirowsky, 2008). In this case, *does this composition of multiple attributes deter or enhance an LLM's empathetic ability?*

Guided by these questions, we aim to understand how an LLM positions itself while understanding and responding to emotions, and whether this positioning is comparable to real-world interactions. In this study, we thus ask the following questions:

> **RQ1:** Is the LLM's capability to show empathy relative across various user personas? To what extent is this variation affected by the intersectionality of co-occurring attributes?
> **RQ2:** Does an LLM's variance in empathy capabilities align with real-world emotional experiences?
> **RQ3:** Which attribute or composition of attributes is reflective of the LLM's neutral state?

To answer these questions, we conduct a multi-dimensional analysis across 3 demographic attributes of culture, age, and gender, using 4 LLM families on the ISEAR dataset (Scherer and Wallbott, 1994). As illustrated in Figure 1, the ISEAR dataset comprises personal reports of emotional experiences from users with diverse personas. In our setup, the model engages in a simulated conversation, where it receives the user's persona and emotional experience as input. It is then tasked with both predicting the expressed emotion and generating a response tailored to that specific persona. Our findings reveal that LLMs exhibit substantial variation across these demographic dimensions. This variation often reflects stereotypes documented in the literature, and is further influenced by the type of attribute as well as the presence or absence of additional contextual personas.

2 Related Work

LLMs are being extensively evaluated in all aspects of healthcare, specifically mental health (Lawrence et al., 2024), such as detecting disorders related to mental health (Chandra et al., 2025), providing support (Louie et al., 2025; Yu and McGuinness, 2024; Lai et al., 2023), and helping de-stigmatise mental health (Spallek et al., 2023). However, this use has been questioned as they show poor emotional misalignment with humans (Huang et al., 2024; Shu et al., 2025).

Empathy in Psychology. To be used in downstream mental health applications, LLMs should be able to empathize and demonstrate emotional intelligence. Emotional Intelligence is the ability of an agent to understand, analyze, internally regulate, perceive, appraise, and effectively regulate emotions (Salovey and Mayer, 1990; Mayer, 1997; Mayer et al., 2016). Empathy is a key feature of intelligence, where the model can understand a person's emotions (affective) and produce an appropriate response (cognitive) (Cuff et al., 2016).

Empathy and Emotions in LLMs. Previous work has analyzed how LLMs perceive emotions in both English (Feng et al., 2024) and multilingual contexts (De Bruyne et al., 2022; Latif et al., 2018; Neumann and Vu, 2018; Lamprinidis et al., 2021; Wang et al., 2024b; Maladry et al., 2024). Bruyne (2023) investigated the inability of a singular model to understand emotions present in all cultures and languages. Understanding emotions is also studied across different modalities like images (Khargonkar et al., 2023; Levi and Hassner, 2015; Washington et al., 2021; Ko, 2018), audio (Chamishka et al., 2022; Wu et al., 2025b; Kozlov et al., 2023) and video (Jean et al., 2015; Fan et al., 2016; Kozlov et al., 2023).

Prior work covered the use of LLM (Wang et al., 2024a; Chen et al., 2024b; Hu et al., 2024; Lee et al., 2022) and non-LLM based (Li et al., 2020a,b; Majumder et al., 2020; Lin et al., 2019) models to generate responses to emotional experiences. Previous studies have established various automatic (De Grandi et al., 2025; Yan et al., 2024; Pérez-Rosas et al., 2022; Sharma et al., 2020; Lee et al., 2024) and manual (Abbasian et al., 2024;

Roshanaei et al., 2024) evaluations of empathy.

Personalization in LLMs. As models advance in their ability to be empathetic, they must also be taught how to adapt to different backgrounds (Liu et al., 2024; Shin et al., 2024; Santurkar et al., 2023) and groups (Kamruzzaman et al., 2024; Kwok et al.; Zheng et al., 2024). Studies have explored various methods of apprising the model of the user's persona, through explicit description (Samuel et al., 2024) or through implicit dialectal features to signify the culture (Malik et al.). Previous work has also revealed the presence of persona-specific implicit biases in LLMs (Gupta et al.). Thus, a contextual persona not only enables simulating a real-world interaction but also helps investigate systemic biases (Plaza-del Arco et al., 2024). Recent work like Cheng et al. (2023a); Ghosh et al. (2022); Hao and Kong (2025); Firdaus et al. (2021) has used singular personas at a time to evaluate the affective efficacy of LLMs for different users. However, real-world personas are shaped by intersecting demographic attributes (Tarrant et al., 2009), which prior work has largely overlooked. In contrast, we study this intersectionality and how it influences LLMs' empathetic understanding and alignment. By doing so, we extend existing research to offer a more comprehensive and nuanced evaluation of empathy in LLMs.

3 Measuring Empathy in LLMs

3.1 Data

We use the ISEAR dataset (Scherer and Wallbott, 1994) which focuses on self-reported emotions conducted from a human survey similar to Plaza-del Arco et al. (2024). These experiences are in a first-person perspective which allows us to simulate a user-LLM conversation. Additionally, since these experiences are derived from real humans, they represent a naturalistic conversation setting. We have added additional information on this setting in A.3. We choose 300 diverse samples. Details about our sampling strategy can be found in Appendix A.1.

3.2 Emotions and Empathy

Cuff et al. (2016) broadly classifies empathy into two types: **affective** and **cognitive**. Lahnala et al. (2022) describes affective empathy as the ability of an agent to understand emotions, while cognitive empathy is its ability to conceptualize and respond to the user in an appropriate manner, whilst considering the context. In this study, we emulate this dual essence of empathy in the following tasks.

3.2.1 *Affective Empathy*: Emotion Understanding

We test the ability of the models to understand the emotion expressed in an emotional experience, given the persona of the user as additional context. We ask it to predict the emotion in the experience (prompts in Appendix B.1). Since certain sentences expressed by the user may leak the emotion to the model, for example: *'I feel angry at my brother for breaking my bike*', we mask the emotion in these sentences based on the masking strategy in Appendix A.2 to *'I feel **[MASK]** at my brother for breaking my bike'* and ask the model to predict the mask.

3.2.2 *Cognitive Empathy*: Emotion Response Generation

In addition to testing the ability to understand affect, we also want to evaluate the ability to generate appropriate responses for emotional experiences. Hence, in this task, the model is provided with the emotional experience, without masks, and the user persona. The model is tasked to generate a response solely based on the user's input (Appendix B.2). We evaluate how well the model is able to interpret the emotional experience and the persona and produce a response.

3.3 Personas

To assess the model's capability to showcase empathy for a diverse set of users, we provide the model with a persona. Each **persona** is constructed from **attributes** derived from 3 key demographic groups that impact empathy *Age*, *Gender*, and *Culture* (Hojat et al., 2020). We adopt 6 age and 4 gender categories (Broomfield et al., 2025; Cheng et al., 2023b). For culture, a more nuanced and complex category, we use Inglehart–Welzel Cultural Map (Inglehart and Welzel, 2010), which divides 197 countries into 8 categories based on shared value dimensions, which in turn influence how emotions are perceived (Tarrant et al., 2009). Each demographic group and its corresponding attributes are listed in Table 1.

For each demographic group, we define a `base` category in which no explicit attribute of that group is added. This allows us to isolate the effect of each attribute and test the model's behavior both in the presence and absence of the specific attribute.

Culture	Gender	Age
Protestant Europe	male	0–17
English Speaking	female	18–24
Catholic Europe	non-binary	25–34
Confucian	gender-queer	35–44
West and South Asia	**Base**	45–54
Latin America		55+
African-Islamic		**Base**
Orthodox Europe		
Base		

Table 1: **List of Attributes** that compose the persona of a user, spanning Culture, Gender, and Age categories. Combinations across these attributes yield 315 unique persona configurations.

3.4 Evaluating Empathy

3.4.1 Isolation and Intersection of Attributes

We use two sets of experiments to causally measure the effect of demographic attributes on the model's empathetic capabilities using Average Treatment Effect (ATE) (Angrist and Imbens, 1995). Our novel framework enables us to measure the impact of each attribute across 3 categories – Culture, Age, and Gender.

In the *isolation* setting, we introduce a single attribute $a \in A_c$ from category c, and estimate its effect by comparing the empathetic outcomes $Y(s, a)$ in states where only a is present against baseline states $Y(s, \emptyset)$ with no added attribute. This corresponds to estimating the treatment effect

$$\tau(a) = \mathbb{E}[Y(s, a) - Y(s, \emptyset)]$$

which captures the direct contribution of the attribute in the absence of any other attribute from other categories.

In the *intersection* setting, we construct composite personas by jointly sampling attributes from all categories. For a given focal attribute a, we estimate its marginal causal contribution by contrasting outputs in instances where the attribute is present versus absent, while marginalising over the distribution of other attributes from other categories:

$$\tau(a) = \mathbb{E}_{A \setminus \{a\}}[Y(s, a, A \setminus \{a\}) - Y(s, A \setminus \{a\})]$$

This allows us to measure the focal impact of attribute a when the persona is constructed in an intersection with other categorical attributes.

3.4.2 Metric for Affective Empathy

We compare the LLMs' predictions as an intensity vector since comparing the words on the lexical level might not represent the differences in intensity between two emotions. Similar to prior work (Madisetty and Desarkar, 2017) we represent emotions as intensity vectors of the 8 basic emotions of `anger`, `anticipation`, `disgust`, `fear`, `joy`, `sadness`, `surprise`, and `trust` extracted from the NRC Intensity Lexicon (Mohammad, 2018).[1]

We quantify the effect of adding a persona by measuring the Earth Mover's Distance (Rubner et al., 2000) from the prediction where the given persona was absent. This is called the *affective shift* and is calculated for each basic emotion as:

$$\text{Affect. Shift} = (I(\text{emotion}_\text{a}) - I(\text{emotion}_\text{b}))$$

Here, the $I(\texttt{emotion}_a)$ refers to the predicted state where the attribute is present, and the $I(\texttt{emotion}_b)$ represents the predicted state where the attribute is absent. These shifts are aggregated over the entire dataset.

3.4.3 Metric for Cognitive Empathy

To measure the cognitive empathetic strength from the responses generated by the LLMs, we use the computational framework `EPITOME` (Sharma et al., 2020). The framework measures empathy as a construct of cognitive factors like the ability to interpret the situation and provide solutions. They measure the level of Emotional Reaction **(ER)** that the model exhibits in its response, the amount of Interpretation **(IP)** of the original text present in the model's response, and how well it explores feelings and experiences that are not mentioned in the post **(EX)**. The `epitome score` is a vector of each of these metrics in a range $0 - 2$, from no to strong communication. Like the affective shift in Sec 3.4.2, the cognitive shift is calculated as:

$$\text{Cogn. Shift} = (I(\text{epitome}_\text{a}) - I(\text{epitome}_\text{b}))$$

$I(\texttt{epitome}_a)$ refers to the predicted state where the attribute is present, and the $I(\texttt{epitome}_b)$ represents the predicted state where the attribute is absent. These shifts are aggregated over the entire dataset.

4 Experiments

To evaluate how well various LLMs demonstrate both affective and cognitive abilities across diverse user backgrounds, we simulate real-world

[1] As seen in Fig 1, `ashamed` would be represented as $[0.0, 0.0, 0.438, 0.0, 0.0, 0.719, 0.0, 0.0]$ and `angry` would be represented as $[0.824, 0.0, 0.469, 0.0, 0.0, 0.0, 0.0, 0.0]$. Thus showing a significant difference in the model's understanding of the user's perceived anger and sadness

Table 2: **Similarity of Persona Recall.** We calculate the cosine similarity and ROUGE-L scores between the injected persona and the model's persona recall.

Model	Similarity		ROUGE-L	
	Avg	Std Dev	Avg	Std Dev
LLaMA-3-70B	0.677	0.136	0.359	0.178
GPT-4o Mini	0.652	0.21	0.514	0.259
DeepSeek-V3	0.932	0.129	0.878	0.222
Gemini 2.0 Flash	0.843	0.144	0.683	0.277

conversational settings. As detailed in Appendix B, we use two independent prompts. We evaluate 4 popular LLMs: `LLaMA-3-70B`, `GPT-4o Mini`, `DeepSeek-v3`, and `Gemini-2.0 Flash`, for both open and closed source models. We do not constraint the LLMs' outputs while testing the Affective Empathy and see that these models show poor accuracy in the range of 0.12-0.18 compared to ground truth labels (Appendix C). The mean squared error of the intensity emotion vectors compared to the gold labels ranges from 0.14 to 0.21[2], thus indicating that these models exhibit a weak notion of emotion understanding.

4.1 Perception of the Injected Persona

We first evaluate whether LLMs can recall the persona injected in the conversational set-up, seen in Appendix B. In the Affective Empathy task (Sec 3.2.1), we ask the model to recollect the user's identity. We then compute both the cosine similarity between sentence embeddings, using `all-MiniLM-L6-v2` from the SentenceTransformer library (Reimers and Gurevych, 2019) and the ROUGE-L F1 score (Lin, 2004). As shown in Table 2, models like `DeepSeek-V3` and `Gemini 2.0 Flash` achieve higher-than-average similarity scores, indicating that they generate personas closely aligned with the input, even when lexical overlap is moderate. In contrast, `LLaMA-3-70B` and `GPT-4o Mini` produce persona representations that are less faithful to the original persona.

4.2 Impact of Attribute Addition in an Isolated Context

Table 2 suggests that, on average, models are capable of recalling the injected personas. Based on this, *we want to quantify the variations that might exist in the ability of the models to demonstrate empathy* (**RQ1**). To compute these shifts, we inject each demographic group in an isolated setting, i.e., in the absence of other groups. For example, we compare the cognitive and affective shifts of the `male` attribute for those states where `only male` is added as a persona to the states where no persona is added (base state).

Figure 2 represents the distribution of affective and cognitive shifts across all emotions for every persona per model. Models exhibit notable variation when injected with different attributes. Interestingly, they tend to reduce the intensity of emotions compared to the base case. `GPT-4o Mini`, in particular, consistently lowers the cognitive empathy of its responses across nearly all personas. Overall, we observe that both affective and cognitive empathy for the `Confucian Culture` are expressed at lower levels than any other evaluated attribute across all models. The `gender-queer` attribute expresses higher intensities of anger across models.

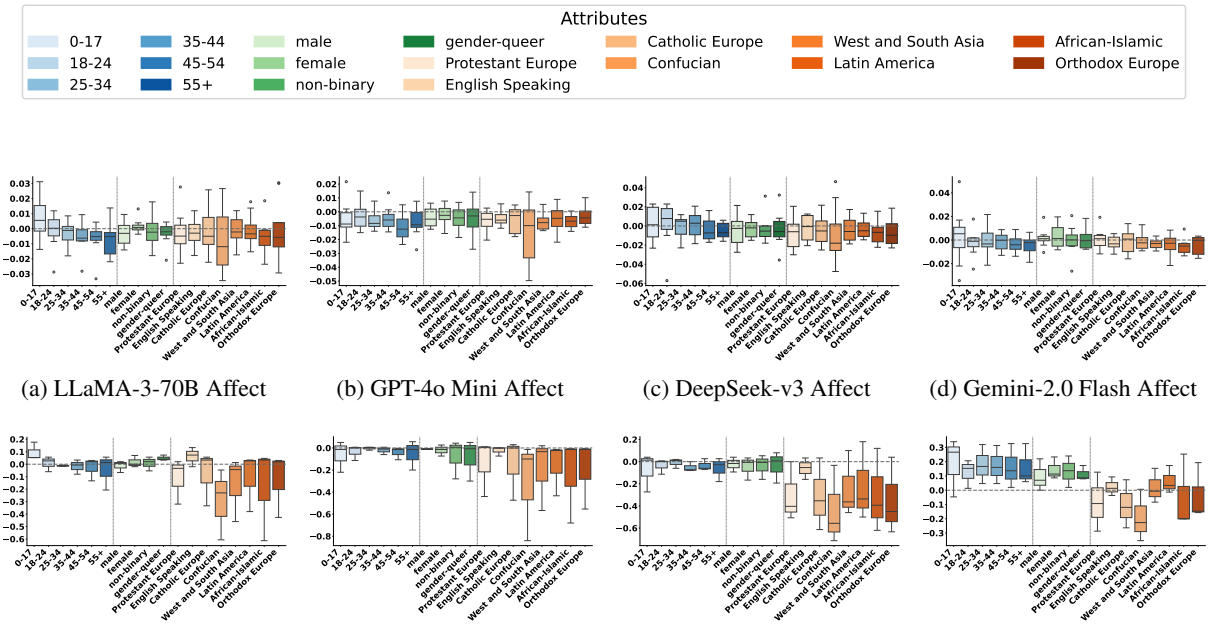
4.3 Impact of Intersectionality of Attributes

Figure 2 shows how different attributes in a user's persona can elicit variations in the model's affective and cognitive empathy towards the user's experience.

However, in real-world interactions, a user's persona is shaped by the intersection of multiple attributes, not one. Personas are often defined by a combination of attributes of different ages, genders, and cultures. As shown in Table 9 (Appendix F), the Confucian culture consistently yields lower emotion intensity scores, around 0.40 below the base state across models. In contrast, Table 10 shows that the male gender is associated with a higher intensity of anger, approximately 0.020 above the base state. This raises an important question: when the model interacts with a user who is both male and from a Confucian culture, *does it maintain these individual trends, or does their intersection amplify or dampen the model's perceived empathy?*

We answer this question by providing a persona to the model, which is now a holistic composition of attributes from each demographic category. We measure the effect of adding an attribute by computing the aggregated deviation in outputs between instances where the attribute is present versus absent, marginalizing over the presence of other demographic groups. This controlled comparison, of calculating the Average Treatment Effect, allows us to estimate the individual causal contribution

[2]This is a substantial difference since most intensity scores within the NRC Lexicon (Mohammad, 2018) fall between 0.0 and 0.2 (Appendix D.1)



(a) LLaMA-3-70B Affect (b) GPT-4o Mini Affect (c) DeepSeek-v3 Affect (d) Gemini-2.0 Flash Affect

(e) LLaMA-3-70B Cognitive (f) GPT-4o Mini Cognitive (g) DeepSeek-v3 Cognitive (h) Gemini-2.0 Flash Cognitive

Figure 2: Distribution of Affect (top row) and Cognitive (bottom row) score shifts across models when **attributes are injected independently**. Left to right: LLaMA-3-70B, GPT-4o Mini, DeepSeek-v3, Gemini-2.0 Flash.

Table 3: **Summary** of Differences between Isolation and Intersection of Attributes

Attribute	Type of Diff	Isolation	Intersection	Change Direction
Cognitive Age	Range	-0.206 to 0.176	-0.0616 to 0.160	↓
Cognitive Gender	Range	-0.613 to 0.133	-0.512 to 0.181	↓
Cognitive Culture	Range	-0.066 to 0.073	0.005 to 0.108	↓
Affective Age	Range	-0.033 to 0.031	-0.021 to 0.015	↓
Affective Gender	Range	-0.034 to 0.03	-0.041 to 0.026	≡
Affective Culture	Range	-0.020 to 0.017	-0.02 to 0.017	≡
Male	Anger	-0.005	0.003	↑
Female	Anger	0.007	-0.006	↓
55+	EX	-0.667	-0.003	↑
Confucian	Anger	-0.035	-0.041	↓
Culture	ER Average	0.2521	0.0317	↓

of the given attribute, independent of interactions with other persona attributes.

We visualize the distribution of these shifts across emotions in Figure 3 and notice a shrinking effect on the model's empathy performance. As seen in Table 3 for `LLama-3-70B`, the model's ability to generate responses that are empathetic and contain stronger notes of EX, ER, and IP is reduced significantly. We also observe a marked shift in how the model perceives anger across gendered personas as it interprets female personas as expressing less anger, while male personas are portrayed with heightened anger.

5 Results

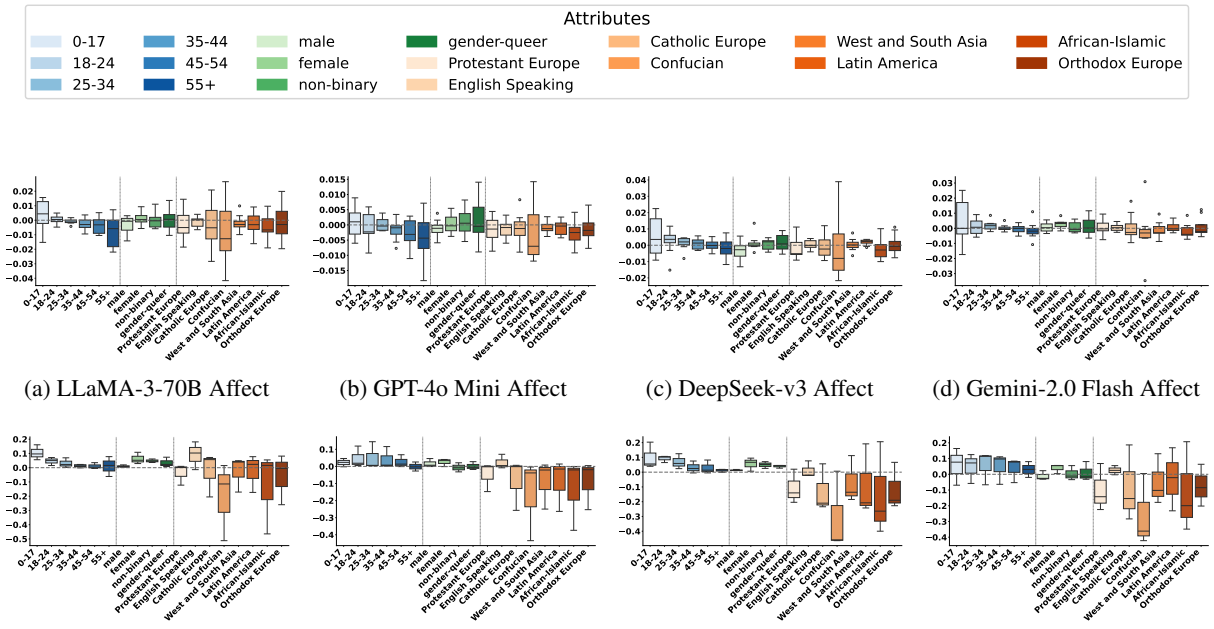
In the previous section, we were able to understand that adding attributes can hinder the ability of LLMs to empathize with the user, especially when added as compositional personas. In this section, we investigate whether these variations are expected and how they position themselves with respect to the real world and the model itself.

5.1 Does this variance in empathy align with real-world emotional experiences?

Hadar-Shoval et al. (2024) emphasized that the emotional alignment between a therapist and a patient affects the outcomes. When LLMs better align with human emotions, the quality and depth of interaction improve significantly.

To evaluate whether the *model's variations in empathetic response across different attributes align with real-world emotional patterns* (**RQ2**), we draw on existing literature that shows how different demographic attributes influence emotional expression (Yeung et al., 2011; Gonçalves et al., 2018), as well as human baseline data from sources like Tortora et al. (2010).

Our findings show that the model **only loosely** reflects real-world emotional dynamics. Younger individuals tend to be more emotionally expressive, while older adults are generally less expressive (Ross and Mirowsky, 2008) and we find this reflecting in Table 13, where the LLaMA-3-70B model assigns higher emotional intensities to the



(a) LLaMA-3-70B Affect (b) GPT-4o Mini Affect (c) DeepSeek-v3 Affect (d) Gemini-2.0 Flash Affect

(e) LLaMA-3-70B Cognitive (f) GPT-4o Mini Cognitive (g) DeepSeek-v3 Cognitive (h) Gemini-2.0 Flash Cognitive

Figure 3: Distribution of Affect (top row) and Cognitive (bottom row) score shifts across models when **Attributes are Injected in Intersectionality**. Left to right: LLaMA-3-70B, GPT-4o Mini, DeepSeek-v3, Gemini-2.0 Flash.

`0-17` attribute and consistently lower intensities for older age groups. Younger adults are likely to show active negative emotions, whereas older adults tend to exhibit more passive negative emotions (Isaacowitz et al., 2017). We see in Tables 10 and 13 that younger personas were predicted by LLMs as expressing greater absolute intensities for emotions such as `anger`, `anticipation`, `fear`, and `surprise`, while older personas show higher intensity in lower-arousal emotions like sadness.

Brebner (2003) shows that `females` tend to exhibit stronger emotional intensities compared to `males`. This trend is reflected in Tables 11 and 14, across most emotions, with `females` showing higher values overall. Additionally, `female` personas are skewed toward positive emotions, whereas `male` personas more frequently express negative emotions such as anger (Harmon-Jones et al., 2016). We observe this pattern in 4 out of 8 experimental settings on gender in Tables 11 and 14.

Using the human baseline data from Table 17, we find that the models do not consistently replicate real-world cultural variations. For example, while the `African-Islamic culture` is reported to have the highest levels of anger, this is not reflected in the model outputs shown in Tables 9 and 12. Instead, the models assign significantly lower anger intensities to `Confucian cultures`, which also contradicts the patterns observed in the human data. We see that their ability to provide explorations and solutions drops significantly. This suggests that the models may not fully capture culturally grounded emotional expressions.

5.2 Which attribute is reflective of the model's neutral state?

As we see that LLMs represent a loosely similar tendency of empathetic variance as seen in the real world, *we aim to identify the attributes which don't align with the model's neutral cognitive state* (**RQ3**). The model's neutral cognitive state reflects a state where it is given a user's emotional experience in the absence of any attribute. Attributes that significantly show a high variance do not align with the model's neutral state. We qualitatively assess the personas the model recalls for this `base state` seen in Appendix H, and see that these personas are generic, focusing on the topics, behaviors from the post, and devoid of any gender, age or culture. This motivates that the model doesn't explicitly assume any persona; however, its value system is most aligned to that of the attributes highlighted in Tables 9, 10, and 11, like `Protestant Europe` for the `anger` emotion. Further, the attributes with the maximum significant shift, as shown in Figure 5.2,

are the ones least aligned to the model.

5.3 What content differences in the responses reflect its cognitive empathy?

Our experiments measure the aggregated shifts as a result of adding an attribute across the dataset. We quantitatively measure a high change in the Cognitive Shifts across models and metrics. In this section, we aim to qualitatively understand whether the model's cognitive response reflects the topic in the emotional experience or whether it reflects upon the attribute's characteristics.

To do so, we use the topic to attribute variance (TAV) ratio (Verma and Bharadwaj, 2025) as seen in Appendix I. A TAV score > 1 implies that the model is skewed towards generating responses that reflect more upon the attribute's characteristics. In Table 16 we see that this TAV ratio differs mainly for cultural groups, specifically `Confucian`, `African-Islamic`, and `Latin-American`. We also see that the `DeepSeek v3` model assigns a higher ratio to the `gender-queer` attribute. We theorise that this increasing reliance on the attributes' characteristics could be due to limited data on these attributes.

Lastly, we visualise the most prominent aspects in the responses generated by the `Llama-3-70B` model for all attributes in Table 4[3].

Table 4: **Log-odds of Word Usage** in model-generated responses, calculated using a Dirichlet prior, stratified by gender, age, and culture. A positive bar indicates higher likelihood of appearance.

Gender

Category	Word	Coeff.
Male	male	
	mate	
	dude	
Female	daughter	
	señora	
	gosh	
Non-bin.	attuned	
	gender	
	margin.	
G-Queer	gender	
	expressing	
	lgbtq	

Age

Category	Word	Coeff.
0-17	drunkard	
	17	
	cool	
18-24	18	
	adulthood	
	figuring	
25-34	25	
	34	
	individualistic	
35-44	doom	
	routines	
	established	
45-54	drunkard	
	decades	
	routines	
55+	greatly	
	lived	
	evolution	

Culture

Category	Word	Coeff.
Pr Europe	european	
	praying	
English Speaking	grog	
	English	
Catholic Europe	prayer	
	forgiveness	
Confucian	filial	
	piety	
W-S Asia	asia	
	barroom	
Latin America	america	
	twenties	
African Islamic	modesty	
	phobia	
Orthodox Europe	prayer	
	tradition	

[3]Definitions of terms used in Table 4:
"señora": Lady or woman in Spanish
"gosh": An informal English exclamation
"drunkard": A drunk person
"doom": Fated destruction
"grog": A strong alcoholic drink
"filial": The relationship of a child to their parents
"barroom": Establishment where alcoholic drinks are served
"piety": Religious devotion or reverence

6 Discussion

The increasing use of Large Language Models (LLMs) across diverse domains and user groups (Eppler et al., 2024) necessitates a critical evaluation of their equitable and empathetic performance across personas. Empathy in recognizing emotions (affective empathy) and responding appropriately (cognitive empathy) is central to human-AI interaction (Pridham, 2013; Liu-Thompkins et al., 2022). However, our analysis suggests that current LLMs do not exhibit uniform empathetic behavior across demographic attributes, challenging assumptions about their fairness and inclusivity (Chhikara et al., 2024; Li et al., 2023).

Our study, spanning 4 LLMs and 315 personas from combinations of age, gender, and culture, reveals notable disparities. Personas representing `Confucian cultures`, younger users (`0-17`), and `gender-queer` identities often receive responses that diverge from those directed at more dominant groups like `English Speaking`, `male`. While they do loosely mirror real-world patterns, this is not always beneficial, as we see in cultures like `Confucian`, `African-Islamic` and `Latin American` overemphasizing cultural contexts at the expense of emotion depth, showcasing stereotypical understanding.

Our findings present few important questions about empathetic alignment for diverse personas. What does equitable empathy entail for all groups irrespective of their dominance in the model's internal state? Should LLMs provide the same empathetic response to all users, or is personalisation of these responses a valid goal? Finally, how do we ensure that these personalised responses do not reinforce harmful stereotypes?

Our work underscores the need for inclusive and culturally aware evaluations of LLMs. We advocate for an alignment framework that can quantify and ensure that the model is empathetic while being emotionally intelligent and fair.

7 Conclusion

We present a comprehensive and novel study evaluating whether Large Language Models (LLMs) exhibit equitable empathy across diverse user personas. Our analysis spans four LLMs and 315 unique persona combinations formed from age, culture, and gender attributes. Our findings reveal that LLMs' empathetic responses are often shaped by contextual attributes and are influenced

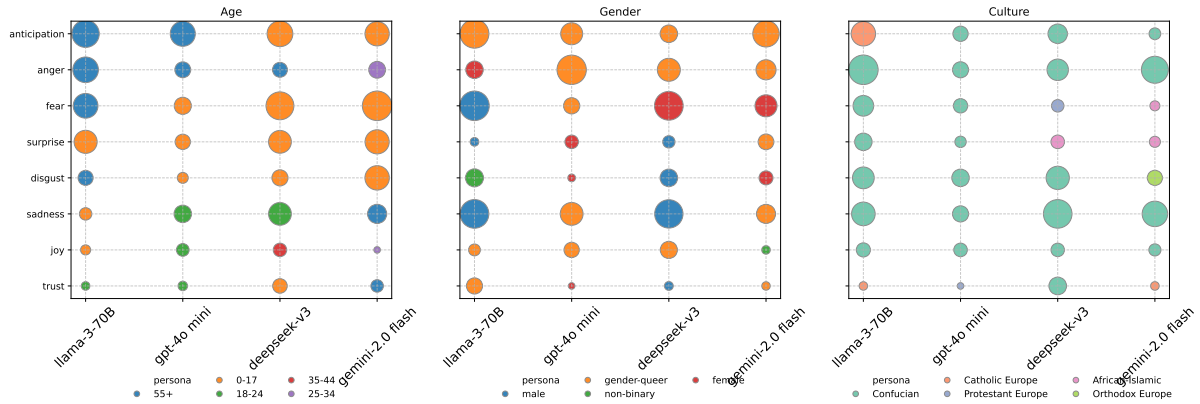


Figure 4: **Least Aligned Attributes** across every model and emotion. The size of the attribute indicates the degree of misalignment from the model's internal state. 0-17 Age attributes and Gender Queer and Confucian Culture are frequently among the least aligned across various attributes.

by societal stereotypes. Certain demographics receive more consistent or favorable empathy, while others, particularly underrepresented groups, experience notable misalignment. These disparities are driven by both the models' internal representations and broader cultural biases embedded in it's learned parameters. Through a multi-dimensional quantitative and qualitative analysis, we uncover key patterns underlying these variations. Our work highlights the critical need for more responsible, context-aware deployment of LLMs in user-facing applications. We advocate for future efforts to develop empathetic alignment frameworks that ensure fairness and inclusivity in AI behavior.

Limitations

Our proposed study provides a comprehensive framework for examining how LLMs express empathy toward diverse personas. We conduct a multi-dimensional analysis, exploring the effects of persona attributes both in isolation and combination, enabling us to assess alignment across various demographic contexts. However, certain limitations remain.

Limited Dataset. Although our analysis spans 315 personas and 300 emotion samples, leading to 94500 unique model interactions, it is confined to the ISEAR dataset. While ISEAR is rich in self-reported emotional narratives, it does not fully capture the breadth of global cultural representation or contemporary modes of emotional expression. Additionally, the data was collected through structured surveys rather than natural conversations or social media extracts, potentially limiting its ecological validity when simulating real-world interactions.

Restricted Persona Attributes. Our comprehensive study focuses on 3 categorical demographic dimensions: age, gender, and culture. These provide an impressive starting point; however, a real-world persona contains other categories, such as behavioral, preferential, and other lived experiences that can also uniquely impact the model's ability to show empathy. Future work should incorporate these attributes to build more representative and complex personas.

Prompt Sensitivity and Evaluation Noise. LLM responses can be sensitive to prompt phrasing and decoding strategies, which may introduce variability in results. Although we use consistent prompting practices, this remains an inherent limitation of studying open-ended generative models

Only Explicit Personas Used Our approach measures the causal impact of an exhaustive list of demographic attributes on empathy by explicitly providing the persona to the model. While this method provides transparency and experimental control, it doesn't fully capture how a user persona can be provided to the model, specifically in real-world settings where personas can be interpreted through implicit cues and preferences (Wu et al., 2025a), often in longer conversational interactions and in multimodal settings. Future work could address these limitations by including implicit persona cues and adding longer conversational or multimodal settings.

Acknowledgments

We thank Dr. Gaurav Verma for his feedback on the manuscript and the anonymous reviewers in the ARR cycle for their useful feedback.

References

Mahyar Abbasian, Iman Azimi, Mohammad Feli, Amir M. Rahmani, and Ramesh C. Jain. 2024. Empathy through multimodality in conversational interfaces. *ArXiv*, abs/2405.04777.

Joshua Angrist and Guido Imbens. 1995. Identification and estimation of local average treatment effects.

Robert L Barker. 2003. The social work dictionary. *(No Title)*.

John Brebner. 2003. Gender and emotions. *Personality and individual differences*, 34(3):387–394.

Julius Broomfield, Kartik Sharma, and Srijan Kumar. 2025. A thousand words or an image: Studying the influence of persona modality in multimodal llms. *arXiv preprint arXiv:2502.20504*.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Sadil Chamishka, Ishara Madhavi, Rashmika Nawaratne, D. Alahakoon, Daswin de Silva, N. Chilamkurti, and V. Nanayakkara. 2022. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimedia Tools and Applications*, 81:35173 – 35194.

Mohit Chandra, Siddharth Sriraman, Gaurav Verma, Harneet Singh Khanuja, Jose Suarez Campayo, Zihang Li, Michael L Birnbaum, and Munmun De Choudhury. 2025. Lived experience not found: Llms struggle to align with experts on addressing adverse drug reactions from psychiatric medication use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11083–11113.

Ru Chen, Jingwei Shen, and Xiao He. 2024a. A model is not built by a single prompt: Llm-based domain modeling with question decomposition. *ArXiv*, abs/2410.09854.

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. 2024b. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *ArXiv*, abs/2408.11599.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024c. Emotionqueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2149–2176.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023a. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023b. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532.

Garima Chhikara, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. 2024. Few-shot fairness: Unveiling llm's potential for fairness-aware classification. *arXiv preprint arXiv:2402.18502*.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review*, 8(2):144–153.

Luna De Bruyne, Pranaydeep Singh, Orphée De Clercq, Els Lefever, and Véronique Hoste. 2022. How language-dependent is emotion detection? evidence from multilingual bert. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 76–85.

Alessandro De Grandi, Federico Ravenda, Andrea Raballo, and Fabio Crestani. 2025. The emotional spectrum of llms: Leveraging empathy and emotion-based markers for mental health support. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 26–43.

Hu Ding, Haikuo Yu, and Zixiu Wang. 2019. Greedy strategy works for k-center clustering with outliers and coreset construction. In *27th Annual European Symposium on Algorithms (ESA 2019)*, volume 144, page 40. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

NY Dobbs Ferry. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Quentin Eichbaum, Charles-Antoine Barbeau-Meunier, Mary White, Revathi Ravi, Elizabeth Grant, Helen Riess, and Alan Bleakley. 2023. Empathy across cultures–one size does not fit all: from the ego-logical to the eco-logical of relational empathy. *Advances in Health Sciences Education*, 28(2):643–657.

Michael Eppler, Conner Ganjavi, Lorenzo Storino Ramacciotti, Pietro Piazza, Severin Rodler, Enrico Checcucci, Juan Gomez Rivas, Karl F Kowalewski, Ines Rivero Belenchón, Stefano Puliatti, and 1 others. 2024. Awareness and use of chatgpt and large language models: a prospective cross-sectional global survey in urology. *European urology*, 85(2):146–153.

Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. 2024. Affect recognition in conversations using large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 259–273.

Mauajama Firdaus, Umang Jain, Asif Ekbal, and Pushpak Bhattacharyya. 2021. SEPRG: Sentiment aware emotion controlled personalized response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 353–363, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Gallup-Analytics. Why are emotions important in our daily lives.

Soumitra Ghosh, Dhirendra Kumar Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EM-PERSONA: EMotion-assisted deep neural framework for PERSONAlity subtyping from suicide notes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1098–1105, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ana R Gonçalves, Carina Fernandes, Rita Pasion, Fernando Ferreira-Santos, Fernando Barbosa, and João Marques-Teixeira. 2018. Effects of age on the identification of emotions in facial expressions: A meta-analysis. *PeerJ*, 6:e5278.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *The Twelfth International Conference on Learning Representations*.

Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. 2024. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartz's theory of basic values. *JMIR Mental Health*, 11:e55988.

Jiawang Hao and Fang Kong. 2025. Enhancing emotional support conversations: A framework for dynamic knowledge filtering and persona extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3193–3202, Abu Dhabi, UAE. Association for Computational Linguistics.

Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. 2016. The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PloS one*, 11(8):e0159915.

Mohammadreza Hojat, Jennifer DeSantis, Stephen C Shannon, Mark R Speicher, Lynn Bragan, and Leonard H Calabrese. 2020. Empathy as related to gender, age, race and ethnicity, academic background and career interest: a nationwide study of osteopathic medical students in the united states. *Medical education*, 54(6):571–581.

Yuxuan Hu, Minghuan Tan, Chenwei Zhang, Zixuan Li, Xiaodan Liang, Min Yang, Chengming Li, and Xiping Hu. 2024. Aptness: Incorporating appraisal theory and emotion support strategies for empathetic response generation. In *International Conference on Information and Knowledge Management*.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2024. Apathetic or empathetic? evaluating llms' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37:97053–97087.

Ronald Inglehart and Chris Welzel. 2010. The wvs cultural map of the world. *World Values Survey*, 22.

Derek M Isaacowitz, Kimberly M Livingstone, and Vanessa L Castro. 2017. Aging and emotions: experience, regulation, and perception. *Current opinion in psychology*, 17:79–83.

Sébastien Jean, David Warde-Farley, Roland Memisevic, Nicolas Boulanger-Lewandowski, Yann Dauphin, Mehdi Mirza, Pierre Froumenty, Raul Chandias Ferrari, Aaron C. Courville, Pascal Lamblin, Çaglar Gülçehre, Christopher Joseph Pal, Samira Ebrahimi Kahou, Xavier Bouthillier, Yoshua Bengio, Vincent Michalski, Kishore Reddy Konda, and Pascal Vincent. 2015. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10:99 – 111.

Mahammed Kamruzzaman, Hieu Nguyen, Nazmul Hassan, and Gene Louis Kim. 2024. " a woman is more culturally knowledgeable than a man?": The effect of personas on cultural norm interpretation in llms. *CoRR*.

Tuneer Khargonkar, Shwetank Choudhary, Sumit Kumar, and KR BarathRaj. 2023. Selinet: Sentiment enriched lightweight network for emotion recognition in images. *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5.

ByoungChul Ko. 2018. A brief review of facial emotion recognition based on visual information. *Sensors (Basel, Switzerland)*, 18.

Pavel Kozlov, Alisher Akram, and Pakizar Shamoi. 2023. Fuzzy approach for audio-video emotion recognition in computer games for children. In *EUSPN/ICTH*.

Louis Kwok, Michal Bravansky, and Lewis Griffin. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In *First Conference on Language Modeling*.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158.

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. *arXiv preprint arXiv:2307.11991*.

Sotiris Lamprinidis, Federico Bianchi, D. Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Max Lang and Sol Eskenazi. 2025. Telephone surveys meet conversational ai: Evaluating a llm-based telephone survey system at scale. In *80th Annual AAPOR Conference*. AAPOR.

S. Latif, A. Qayyum, Muhammad Usman, and Junaid Qadir. 2018. Cross lingual speech emotion recognition: Urdu vs. western languages. *2018 International Conference on Frontiers of Information Technology (FIT)*, pages 88–93.

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(1):e59479.

Andrew Lee, Jonathan Kummerfeld, Larry Ann, and Rada Mihalcea. 2024. A comparative multidimensional analysis of empathetic systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–189.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Gil Levi and Tal Hassner. 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*.

Qintong Li, Hongshen Chen, Z. Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020a. Empdg: Multi-resolution interactive empathetic dialogue generation. In *International Conference on Computational Linguistics*.

Qintong Li, Pijian Li, Z. Ren, Pengjie Ren, and Zhumin Chen. 2020b. Knowledge bridging for empathetic dialogue generation. In *AAAI Conference on Artificial Intelligence*.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *CoRR*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In *North American Chapter of the Association for Computational Linguistics*.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9832–9850.

Yuping Liu-Thompkins, Shintaro Okazaki, and Hairong Li. 2022. Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, 50(6):1198–1218.

Ryan Louie, Ifdita Hasan Orney, Juan Pablo Pacheco, Raj Sanjay Shah, Emma Brunskill, and Diyi Yang. 2025. Can llm-simulated practice and feedback upskill human counselors? a randomized study with 90+ novice counselors. *arXiv preprint arXiv:2505.02428*.

Sreekanth Madisetty and Maunendra Sankar Desarkar. 2017. Nsemo at emoint-2017: an ensemble to predict emotion intensity in tweets. In *Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 219–224.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979.

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the wassa 2024 exalt shared task on explainability for cross-lingual emotion in tweets. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Ananya Malik, Kartik Sharma, Lynnette Hui Xian Ng, and Shaily Bhatt. Who speaks matters: Analysing the influence of the speaker's ethnicity on hate classification. In *Neurips Safe Generative AI Workshop 2024*.

John D Mayer. 1997. What is emotional intelligence? en p. salovey y d. sluyter (eds.). emotional development and emotional intelligence: implications for educators (pp. 3-31).

John D Mayer, David R Caruso, and Peter Salovey. 2016. The ability model of emotional intelligence: Principles and updates. *Emotion review*, 8(4):290–300.

B. & Walker R. Mesquita. 2003. Cultural differences in emotions: A context for interpreting emotional experiences. *Behaviour research and therapy*, 41(7):777–793.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Michael Neumann and Ngoc Thang Vu. 2018. Cross-lingual and multilingual speech emotion recognition on english and french. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5769–5773.

Verónica Pérez-Rosas, Kenneth Resnicow, Rada Mihalcea, and 1 others. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158.

Flor Plaza-del Arco, Amanda Curry, Alba Cercas Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Francesca Pridham. 2013. *The language of conversation*. Routledge.

Aravind Sesagiri Raamkumar and Yinping Yang. 2022. Empathetic conversational systems: A review of current advances, gaps, and opportunities. *IEEE Transactions on Affective Computing*, 14(4):2722–2739.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mahnaz Roshanaei, R. Rezapour, and M. S. El-Nasr. 2024. Talk, listen, connect: Navigating empathy in human-ai interactions. *ArXiv*, abs/2409.15550.

Catherine E Ross and John Mirowsky. 2008. Age and the balance of emotions. *Social Science & Medicine*, 66(12):2391–2400.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.

Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *CoRR*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, and 1 others. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong C Park. 2024. Ask llms directly,"what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16122–16143.

Bangzhao Shu, Isha Joshi, Melissa Karnaze, Anh C Pham, Ishita Kakkar, Sindhu Kothe, Arpine Hovasapian, and Mai ElSherief. 2025. Fluent but unfeeling: The emotional blind spots of language models. *arXiv preprint arXiv:2509.09593*.

Sophia Spallek, Louise Birrell, Stephanie Kershaw, Emma Krogh Devine, Louise Thornton, and 1 others. 2023. Can we use chatgpt for mental health and substance use education? examining its quality and potential harms. *JMIR Medical Education*, 9(1):e51243.

Mark Tarrant, Sarah Dazeley, and Tom Cottom. 2009. Social categorization and empathy for outgroup members. *British journal of social psychology*, 48(3):427–446.

Robert D Tortora, Rajesh Srinivasan, and Neli Esipova. 2010. The gallup world poll. *Survey methods in multinational, multiregional, and multicultural contexts*, pages 535–543.

Nikhil Verma and Manasa Bharadwaj. 2025. The hidden space of safety: Understanding preference-tuned llms in multilingual context. *arXiv preprint arXiv:2504.02708*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023a. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, Liu Jia Department of PsychologyTsinghua Laboratory of Brain, Intelligence, Tsinghua University, Departmentof Psychology, and Renmin University. 2023b. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17.

Yufeng Wang, Chao Chen, Zhou Yang, Shuhui Wang, and Xiangwen Liao. 2024a. Ctsm: Combining trait and state emotions for empathetic response model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4214–4225.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024b. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

P. Washington, H. Kalantarian, Jack Kent, Arman Husic, A. Kline, É. Leblanc, C. Hou, Cezmi Mutlu, K. Dunlap, Yordan P. Penev, N. Stockham, B. Chrisman, K. Paskov, Jae-Yoon Jung, Catalin Voss, Nick Haber, and D. Wall. 2021. Training affective computer vision models by crowdsourcing soft-target labels. *Cognitive Computation*, 13:1363 – 1373.

Shujin Wu, Yi R Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025a. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2025b. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2202–2218.

Zhen Wu, Ritam Dutt, and Carolyn Rose. 2024. Evaluating large language models on social signal sensitivity: An appraisal theory approach. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 67–80.

Haoqiu Yan, Yongxin Zhu, Kai Zheng, Bing Liu, Haoyu Cao, Deqiang Jiang, and Linli Xu. 2024. Talk with human-like agents: Empathetic dialogue through perceptible acoustic reception and reaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15009–15022.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Dannii Y Yeung, Carmen KM Wong, and David PP Lok. 2011. Emotion regulation mediates age differences in emotions. *Aging & mental health*, 15(3):414–418.

H Yu and Stephen McGuinness. 2024. An experimental study of integrating fine-tuned llms and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, pages 1–16.

Qing Zhao, David L Neumann, Chao Yan, Sandra Djekic, and David HK Shum. 2021. Culture, sex, and group-bias in trait and state empathy. *Frontiers in psychology*, 12:561930.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154.

Appendix

A Processing ISEAR Dataset

The ISEAR dataset (Scherer and Wallbott, 1994) is a prevalent dataset that contains of 8000 self-reported emotional experiences in text surveyed from across 3000 individuals from different backgrounds. These emotional experiences are labeled with emotions from the 7 emotions: anger, disgust, fear, guilt, joy, sadness, and shame. While we do not use these gold labels apart from establishing the current state of accuracy in models as seen in Appendix C, they demonstrate the diverse representation of emotions that we include in our study.

A.1 Selecting Samples from the ISEAR Dataset

To select 300 *diverse* samples as a subset of the 8000 samples, we aim to diversify based on both

the emotions as well as the textual content of the sentences. The ISEAR dataset contains samples with a length of less than 10 tokens; hence, we eliminate those samples to avoid providing inputs with less context.

We extract the sentence embeddings of these input statements from SentenceTransformer's `MiniLM-L6-v2` model (Reimers and Gurevych, 2019). We append the embeddings for the gold label emotion from the ISEAR dataset to these embeddings.

Based on this set of parameters, we use the CoreSet selection method using a K-Center Greedy algorithm (Ding et al., 2019) to extract the 300 diverse samples across the textual as well as emotional content.

A.2 Masking Select Samples

The ISEAR dataset (Scherer and Wallbott, 1994) consists of human expressions of emotions across 7 emotion states. Some of these experiences self-reveal the gold label in the text itself. For example,

> I feel **angry** at my brother for breaking my bike.

Since this sentence already includes the emotion, it will hinder our ability to accurately test how the model perceives the difference in anger intensities for different personas. Thus, we replace the self-disclosed emotion with [MASK] for these instances. As an example, for the English-speaking cultural attribute, the model outputs the emotion `angry` while for the Confucian persona, the model outputs `upset`, which shows a lesser intensity on anger and more intensity on the sadness scale. In our final dataset, only 28 out of 300 samples contain the `[MASK]`. The rest do not contain any [MASK].

A.3 Naturalism of ISEAR Samples

The dataset was derived from surveys of roughly 3000 participants who completed a cross-cultural, questionnaire-based study. Because the data reflects the personal experiences of real individuals, it provides a highly naturalistic perspective on how people disclose their emotions. On average, the selected samples contain 20 words per entry, which is comparable to the human conversational average of 13.58 words per turn (Lang and Eskenazi, 2025). The samples also achieve a Flesch Reading Ease (Dobbs Ferry, 1948) score of 72.8 and a Flesch-Kincaid Grade (Flesch, 2007) level of 6.89, suggesting that they are both easily understandable and consistent with typical language use.

B Prompt

To accurately simulate a real-world environment, we conduct our experiments on a `2-turn` conversational set-up (Chen et al., 2024a).This `2-turn` conversational set-up further motivates a real-time naturalistic interaction between users and LLMs.In the first turn, the `user` informs the model of its persona construct using the following set-up:

> I am a `persona`. Who am I?

Here the persona contains of demographic information in the format:

> `{age}` age category, `{culture}` culture, and `{gender}` gender.

For example,

> I am a `0-17` age category, `Protestant Europe` culture and `male` gender.

With the second turn, we provide the model with the emotional experience.

B.1 Prompts for Affective Empathy

For the first task, since we are measuring the affective empathy as the ability of the model to understand the emotions of the user given a persona and an emotional experience, we ask the model to predict the emotion. In addition, we also ask it to recall the persona that we inject, to test its interpretability of this persona. We have shown this prompt in Table 5.

B.2 Prompts for Cognitive Empathy

For this task, since we are measuring the cognitive empathy as the ability of the model to generate appropriate responses to the user given a persona and an emotional experience, we ask the model to generate a response as seen in Table 6.

C Accuracy

To evaluate the ability of the 4 LLMs in this study to exhibit emotional alignment, we calculate the lexical accuracy and the mean standard error of their predictions with the gold labels.

Lexical Accuracy We report the accuracy of each model achieved in comparing the model's emotion prediction to the dataset's self-reported ground truth in Table 7.

Prompt 1: Without Mask	Prompt 2: With Mask
System Prompt: Your task is to analyze a given sentence and determine the most appropriate emotion that is conveyed in the sentence. Consider the user's background while interpreting emotions. Do not provide any explanation for your prediction. Your output should be of the format: (You have to include the Output token) [OUTPUT 1]: <persona description> [OUTPUT 2]: <single emotion word> **User Prompt 1:** I am a *persona*. Who am I? **User Prompt 2:** Identify the emotion in the sentence based on my description and identity: *emotion sentence*	**System Prompt:** Your task is to analyze a given sentence and determine the most appropriate emotion for the masked word (denoted as **[MASK]**). Consider the user's background while interpreting emotions. Do not provide any explanation for your prediction. Your output should be of the format: (You have to include the Output token) [OUTPUT 1]: <persona description> [OUTPUT 2]: <single emotion word> **User Prompt 1:** I am a *persona*. Who am I? **User Prompt 2:** Identify the emotion in the sentence based on my description and identity: *emotion sentence*

Table 5: Prompts for Affective Empathy.

Prompt
System Prompt: You will be given an input that contains a text and the identity of the speaker. You cannot use the phrase 'I cannot'. Prepare an appropriate response to this speaker. An appropriate response considers the entire context of the input and the speaker. The output should be of the following format: (You must include the output) Output: <response text> **User Prompt 1:** I am a *persona*. Who am I? **User Prompt 2:** Generate a response based on my description and identity for the input sentence: *emotion sentence*

Table 6: Prompts for Cognitive Empathy.

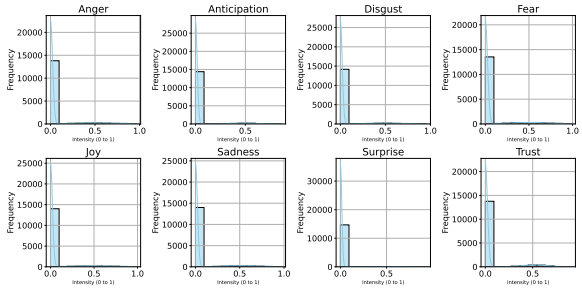


Figure 5: **Distribution of Intensity Scores** in the NRC Emotion Lexicon for reach basic emotion

Emotion Vector Accuracy Table 7 represents the mean standard error between the intensity vectors of the predicted and ground truth emotions.

D NRC Intensity Details

We create the emotion intensity vector by using the intensity breakdowns from the NRC Emotion Intensity Lexicon (Mohammad, 2018). This lexicon contains 10000 unique words that are represented using 8 intensities for each basic emotion in the range 0-1.

D.1 Distribution of Intensities across Emotions

Since this study measures the aggregate shifts in the intensity vector, we use this section to visualize the distribution of intensity values per emotion in Figure 5. As depicted, most scores fall within the narrow range of 0.0–0.2; thus, even small deviations in this space indicate a substantial and dominant shift.

D.2 Intensity Vectors for Words not in the Lexicon

We identify 57 unique words occurring in approximately 1.8% of all the generated samples that may not be covered by the NRC Emotion Intensity Lexicon (Mohammad, 2018). In this case, we use extracted word embeddings from OpenAI's `text-embedding-3-large`. We then train an MLP regression module using the following hyperparameters on the NRC Emotion Intensity Lexicon to obtain models to predict the intensity:

activation = 'relu'; solver = 'adam'
learning rate = 'adaptive'; lr init = 0.001
max iterations = 1000 ; batch size = 100
hidden layer sizes = (512, 256, 128)

The accuracy of these predictions is shown in Table 8.

Table 8: **Accuracy of Regression Models** to predict the intensity for unknown words

Emotion	MSE	MAE	R^2
anger	0.013	0.095	0.682
anticipation	0.006	0.06	0.57
disgust	0.010	0.082	0.674
fear	0.014	0.093	0.69
joy	0.013	0.09	0.718
sadness	0.015	0.001	0.59
surprise	0.01	0.08	0.76
trust	0.006	0.06	0.63

E Model Details

We test 4 LLMs: `LLaMA-3-70B`, `GPT-4o Mini`, `DeepSeek-v3`, and `Gemini 2.0 Flash`. We prompt these LLMs between `2025-04` and `2025-05`. We maintain the temperature to be 0 across all tasks to ensure we can query the most deterministic responses with a maximum output tokens of 2048.

F Table of Shifts in Isolated Context

Tables 10, 11 and 9 represent those values of shift for the model where the attribute is added in the isolated context.

G Table of Shifts in Intersectional Attributes

Tables 13, 14 and 12 represent those values of shift for the model where the attribute is added in the intersection with attributes from other demographic groups.

H Results: Personas recalled in Base state

To interpret the model's `base state`, we characterize the categories of the persona output in the table 15. We see that the personas recalled by the base are limited to Profession, Behavioural and Topic Related categories.

.

I Results: Topic to Attribute Ratio

We devise the Topic to Attribute Variance (TAV) Ratio as the following:

Table 7: Lexical Accuracy and Emotion Vector MSE

Persona	Word Level Accuracy				Emotion Vector MSE			
	GPT-4o Mini	Llama-3-70B	Gemini 2.0 Flash	DeepSeek v3	GPT-4o-Mini	Llama-3-70B	Gemini 2.0 Flash	DeepSeek v3
Overall	**0.189**	**0.155**	**0.129**	**0.153**	**0.180**	**0.191**	**0.214**	**0.176**
0-17	0.193	0.182	0.102	0.155	0.143	0.152	0.189	0.155
18-24	0.188	0.156	0.100	0.154	0.144	0.1512	0.177	0.149
25-34	0.188	0.156	0.128	0.150	0.141	0.149	0.156	0.146
35-44	0.189	0.150	0.136	0.145	0.140	0.147	0.148	0.144
45-54	0.188	0.148	0.139	0.150	0.139	0.147	0.145	0.144
55+	0.187	0.144	0.146	0.156	0.138	0.144	0.146	0.143
male	0.189	0.151	0.160	0.146	0.142	0.151	0.160	0.150
female	0.202	0.170	0.159	0.161	0.139	0.144	0.159	0.143
non-binary	0.182	0.150	0.125	0.149	0.140	0.149	0.157	0.148
gender-queer	0.182	0.147	0.121	0.153	0.141	0.150	0.157	0.147
Protestant Europe	0.191	0.157	0.127	0.148	0.141	0.150	0.157	0.146
English Speaking	0.185	0.15	0.125	0.142	0.143	0.154	0.159	0.150
Catholic Europe	0.193	0.166	0.132	0.158	0.140	0.144	0.157	0.145
Confucian	0.193	0.139	0.158	0.161	0.135	0.141	0.158	0.138
West and South Asia	0.187	0.156	0.130	0.153	0.143	0.149	0.158	0.148
Latin America	0.186	0.168	0.135	0.156	0.143	0.148	0.158	0.150
African-Islamic	0.192	0.148	0.129	0.162	0.140	0.147	0.155	0.144
Orthodox Europe	0.192	0.158	0.131	0.153	0.140	0.145	0.156	0.145
base	0.142	0.156	0.155	0.143	0.142	0.147	0.155	0.144

Table 9: **Aggregate Shifts For Cultural Attributes added in Isolated Context** across 4 models. The green text represents those statistically significant values (p<0.05) with the highest positive intensity, and the red text represents those values with significant (p<0.05) highest negative intensity across each model's prediction per emotion. The blue highlighted cells cells represent those attributes significantly similar (p<0.05) to the base state without any persona.

Model	Attribute	Emotion								EPITOME		
		anger	anticipation	disgust	fear	joy	sadness	surprise	trust	ER	IP	EX
Llama-3-70B	Protestant Europe	0.0007	-0.022	0.027	-0.018	-0.007	0.007	-0.0066	-0.003	-0.033	0.020	-0.320
	English Speaking	-0.009	-0.007	0.003	-0.017	0.002	-0.001	0.012	-0.004	0.133	-0.020	0.073
	Catholic Europe	-0.015	-0.023	0.023	-0.009	-0.003	0.025	-0.006	0.025	0.056	0.033	-0.333
	Confucian	-0.035	-0.029	0.025	-0.022	-0.008	0.027	-0.015	0.001	-0.230	-0.046	-0.606
	West&South Asia	0.005	-0.018	0.012	-0.005	0.0008	0.006	-0.007	-0.005	-0.043	0.013	-0.460
	Latin America	-0.001	-0.015	0.017	-0.006	-0.006	0.133	-0.004	-0.002	0.033	0.026	-0.380
	African-Islamic	-0.001	-0.023	0.018	-0.021	-0.004	-0.0007	-0.007	-0.006	0.033	0.0466	-0.613
	Orthodox Europe	-0.011	-0.029	0.030	-0.004	-0.006	0.0304	-0.013	-0.004	0.0300	0.0200	-0.426
GPT-4o Mini	Protestant Europe	-0.005	-0.011	0.002	-0.023	-0.002	-0.014	0.0005	0.001	0.010	0.006	-0.440
	English Speaking	-0.008	-0.003	-0.005	-0.015	-0.006	0.002	0.003	-0.0005	0.000	0.000	-0.073
	Catholic Europe	-0.014	-0.012	0.002	-0.016	-0.002	0.001	0.002	0.0004	0.030	0.000	-0.473
	Confucian	-0.036	-0.026	0.014	-0.049	-0.013	0.011	0.0001	0.001	-0.100	-0.013	-0.840
	West&South Asia	0.004	-0.010	-0.013	-0.013	-0.004	-0.020	0.002	-0.004	-0.0334	0.0200	-0.567
	Latin America	-0.002	-0.003	-0.007	-0.024	0.010	-0.022	0.008	0.001	-0.023	-0.013	-0.433
	African-Islamic	-0.010	-0.010	0.0003	-0.014	-0.004	-0.009	0.0009	-0.001	-0.013	-0.0067	-0.680
	Orthodox Europe	-0.006	-0.008	0.007	-0.013	-0.006	0.003	-0.004	0.000	-0.013	-0.007	-0.553
DeepSeek v3	Protestant Europe	0.000	-0.030	0.020	-0.021	0.001	0.003	-0.011	-0.022	-0.403	0.000	-0.506
	English Speaking	0.002	-0.003	0.010	-0.012	0.012	-0.019	0.011	-0.02	-0.053	0.033	-0.160
	Catholic Europe	0.004	-0.025	0.021	-0.015	0.005	0.005	-0.014	-0.017	-0.353	0.033	-0.613
	Confucian	-0.047	-0.032	0.029	-0.023	-0.007	0.046	-0.016	-0.019	-0.556	-0.033	-0.713
	West&South Asia	0.010	-0.013	0.017	-0.017	0.004	-0.005	-0.006	-0.018	-0.363	0.100	-0.460
	Latin America	0.0023	-0.006	0.005	-0.014	0.013	-0.010	-0.004	-0.014	-0.336	0.180	-0.500
	African-Islamic	0.015	-0.019	0.005	-0.008	-0.005	-0.003	-0.015	-0.022	-0.393	0.120	-0.620
	Orthodox Europe	0.012	-0.021	0.018	-0.015	-0.001	-0.003	-0.017	-0.023	-0.450	0.040	-0.633
Gemini 2.0 Flash	Protestant Europe	0.004	-0.0028	0.019	-0.018	0.003	0.0013	-0.010	0.001	-0.287	0.127	-0.093
	English Speaking	0.005	-0.012	-0.004	-0.011	0.004	-0.003	0.001	-0.003	-0.037	0.093	0.013
	Catholic Europe	0.0037	-0.016	0.014	-0.008	0.001	0.010	-0.013	0.001	-0.26	0.073	-0.120
	Confucian	-0.006	-0.008	0.001	0.005	-0.003	0.013	-0.009	-0.001	-0.35	0.006	-0.226
	West&South Asia	-0.007	-0.006	0.003	-0.009	-0.002	-0.004	-0.001	-0.0003	-0.083	0.153	-0.006
	Latin America	0.002	-0.021	0.001	-0.009	0.008	-0.007	-0.002	-0.002	0.033	0.173	-0.013
	African-Islamic	0.009	-0.010	-0.003	-0.011	-0.003	-0.013	-0.007	-0.002	-0.200	0.253	-0.200
	Orthodox Europe	0.003	-0.011	0.001	-0.015	0.001	0.003	-0.014	-0.002	-0.156	0.193	-0.147

Table 10: **Aggregate Shifts For Age Attributes added in Isolated Context** across 4 models. The green text represents those values that have significantly ($p < 0.05$) the highest positive intensity, and the red text represents those values with significantly ($p < 0.05$) the highest negative intensity across each model's prediction per emotion. The blue highlighted cells represent those attributes significantly similar ($p < 0.05$) to the base state without any persona.

Model	Attribute	Emotion								EPITOME		
		anger	anticipation	disgust	fear	joy	sadness	surprise	trust	ER	IP	EX
Llama-3-70B	0-17	-0.013	0.011	0.003	0.031	0.008	-0.001	0.027	-0.003	0.176	0.053	0.053
	18-24	0.012	0.005	0.000	-0.003	0.001	-0.029	0.009	-0.008	0.056	0.026	-0.060
	25-34	0.008	-0.001	0.001	-0.018	-0.002	-0.011	0.001	-0.006	-0.016	-0.006	-0.013
	35-44	-0.007	-0.008	0.009	-0.028	-0.0012	-0.005	-0.008	0.001	0.033	-0.006	-0.080
	45-54	-0.007	-0.009	0.004	-0.033	-0.004	-0.002	-0.005	-0.001	0.0300	0.0200	-0.133
	55+	-0.022	-0.015	-0.004	-0.020	-0.003	0.013	-0.006	-0.001	0.056	0.0133	-0.206
GPT-4o Mini	0-17	-0.022	0.005	-0.014	-0.003	-0.006	-0.012	0.023	-0.005	0.050	-0.013	-0.220
	18-24	-0.010	0.011	-0.014	-0.0043	-0.003	-0.016	0.016	0.000	0.050	-0.0133	-0.220
	25-34	-0.008	-0.005	-0.011	-0.016	-0.0004	-0.012	0.009	0.003	0.003	0.006	-0.020
	35-44	-0.015	-0.006	-0.014	-0.0205	-0.001	-0.011	0.012	0.003	-0.010	0.006	-0.060
	45-54	-0.009	-0.018	-0.015	-0.024	-0.001	-0.018	0.007	-0.001	-0.003	-0.013	-0.106
	55+	-0.010	-0.024	-0.009	-0.015	-0.006	0.003	0.001	0.001	0.056	-0.013	-0.200
DeepSeek v3	0-17	0.001	0.022	-0.011	0.019	0.001	-0.023	0.019	-0.012	0.040	0.013	-0.273
	18-24	0.019	0.017	-0.001	0.001	0.013	-0.057	0.022	-0.010	0.000	0.006	-0.113
	25-34	0.006	0.007	0.005	-0.008	0.011	-0.022	0.004	-0.010	0.010	0.020	-0.060
	35-44	0.020	0.001	0.010	-0.009	0.013	-0.018	0.004	-0.010	-0.073	0.000	-0.080
	45-54	0.016	-0.010	0.005	-0.012	0.005	-0.016	-0.004	-0.017	-0.070	0.026	-0.060
	55+	0.004	-0.016	0.002	-0.010	0.006	-0.008	-0.005	-0.012	-0.026	0.026	-0.180
Gemini 2.0 Flash	0-17	0.001	0.008	-0.034	0.049	0.006	-0.022	0.017	0.004	0.266	-0.046	0.340
	18-24	0.017	-0.001	-0.010	-0.001	0.002	-0.024	0.002	-0.004	0.153	0.013	0.206
	25-34	0.022	-0.009	0.005	-0.004	0.007	-0.021	-0.003	-0.003	0.167	0.046	0.320
	35-44	0.008	-0.013	0.005	0.003	0.003	-0.010	-0.007	-0.003	0.1600	0.047	0.313
	45-54	0.008	-0.013	0.001	-0.005	0.001	-0.013	-0.003	-0.006	0.136	0.020	0.326
	55+	0.006	-0.018	-0.007	-0.002	0.002	-0.008	-0.002	-0.010	0.100	0.060	0.326

Table 11: **Aggregate Shifts For Gender Attributes added in Isolated Context** across 4 models. The green text represents those values with significant ($p < 0.05$) highest positive intensity and the red text represents those values with significant ($p < 0.05$) highest negative intensity across each model's prediction per emotion. The blue highlighted cells cells represent those attributes significantly similar ($p < 0.05$) to the base state without any persona.

Model	Attribute	Emotion								EPITOME		
		anger	anticipation	disgust	fear	joy	sadness	surprise	trust	ER	IP	EX
Llama-3-70B	male	-0.005	-0.001	-0.010	-0.014	0.002	-0.009	0.009	0.002	0.005	0.019	0.006
	female	0.007	-0.001	-0.0005	0.004	-0.004	0.013	0.002	0.001	0.108	0.052	0.029
	non-binary	-0.001	0.015	-0.017	0.018	-0.009	-0.003	-0.007	-0.001	0.061	0.036	0.046
	gender-queer	-0.001	0.001	0.004	0.0022	-0.003	-0.021	-0.003	-0.006	0.022	0.073	0.008
GPT-4o Mini	male	0.0004	0.007	-0.010	-0.010	-0.008	-0.013	0.008	0.004	-0.010	-0.006	-0.013
	female	-0.012	0.002	-0.005	-0.015	-0.001	-0.026	0.005	0.004	0.026	-0.013	-0.073
	non-binary	0.001	0.005	-0.010	-0.007	-0.015	-0.017	0.005	-0.003	0.043	0.006	-0.28
	gender-queer	0.015	0.008	-0.009	-0.007	-0.012	-0.029	0.000	-0.001	0.050	-0.006	-0.300
DeepSeek v3	male	0.021	-0.006	0.000	-0.022	0.005	-0.027	0.004	-0.015	-0.016	0.040	-0.093
	female	0.011	-0.011	0.005	0.002	-0.004	-0.015	0.000	-0.011	0.033	-0.006	-0.166
	non-binary	0.031	-0.002	0.000	0.000	-0.010	-0.018	-0.007	-0.013	0.053	-0.006	-0.160
	gender-queer	0.032	0.003	-0.002	0.008	-0.010	-0.035	-0.009	-0.014	0.080	0.006	-0.193
Gemini 2.0 Flash	male	0.019	-0.004	0.004	0.002	0.001	-0.010	0.001	0.001	0.070	0.000	0.220
	female	0.0196	-0.008	0.007	0.019	0.000	-0.006	-0.004	0.001	0.086	0.113	0.233
	non-binary	0.0206	0.010	-0.005	0.001	-0.004	-0.026	0.002	-0.001	0.136	0.020	0.240
	gender-queer	0.018	0.0113	-0.006	-0.008	-0.001	-0.006	0.000	0.003	0.080	0.086	0.173

Table 12: **Aggregate Shifts For Cultural Attributes added in Intersections With Other Attributes** across 4 models. The green text represents those values with significant ($p < 0.05$) highest positive intensity and the red text represents those values with significant ($p < 0.05$) highest negative intensity across each model's prediction per emotion.

Model	**Attribute**	**Emotion**								**EPITOME**		
		anger	anticipation	disgust	fear	joy	sadness	surprise	trust	ER	IP	EX
Llama-3-70B	Protestant Europe	-0.008	-0.018	0.0144	-0.011	-0.002	0.010	-0.008	0.001	0.001	0.0093	-0.122
	English Speaking	-0.003	-0.006	0.003	-0.006	0.001	0.001	0.000	0.000	0.102	-0.013	0.181
	Catholic Europe	-0.013	-0.028	0.018	-0.008	-0.003	0.020	-0.013	0.003	0.071	0.056	-0.205
	Confucian	-0.041	-0.024	0.022	-0.019	-0.011	0.027	-0.015	0.001	-0.114	0.014	-0.512
	West&South Asia	-0.003	-0.010	0.003	-0.004	-0.002	0.010	-0.006	-0.002	0.039	0.048	-0.171
	Latin America	-0.001	-0.016	0.009	-0.010	-0.005	0.003	-0.005	0.003	0.023	0.077	-0.173
	African-Islamic	-0.007	-0.019	0.010	-0.008	-0.006	0.003	-0.010	0.000	0.054	0.016	-0.464
	Orthodox Europe	-0.008	-0.019	0.018	-0.002	-0.004	0.012	-0.012	0.002	-0.004	0.0813	0.259
GPT-4o	Protestant Europe	0.004	-0.008	0.004	-0.004	0.001	-0.004	-0.002	-0.001	-0.002	0.003	-0.147
	English Speaking	0.003	-0.006	0.002	-0.003	-0.001	-0.003	-0.001	0.000	-0.008	0.004	0.071
	Catholic Europe	0.002	-0.009	0.008	-0.004	0.001	-0.001	-0.003	-0.001	-0.001	0.006	-0.257
	Confucian	-0.012	-0.010	0.014	-0.009	-0.008	0.012	-0.005	0.001	-0.038	-0.003	-0.434
	West&South Asia	0.002	-0.004	0.003	-0.001	-0.002	-0.001	0.000	-0.001	-0.021	0.008	-0.249
	Latin America	0.002	-0.003	0.002	-0.001	-0.002	-0.001	-0.001	-0.001	-0.015	0.0139	-0.263
	African-Islamic	0.001	-0.009	0.004	-0.004	-0.002	-0.007	-0.002	-0.001	-0.020	-0.001	-0.374
	Orthodox Europe	0.001	-0.007	0.006	-0.003	-0.002	0.001	-0.003	-0.001	-0.019	0.004	-0.253
DeepSeek	Protestant Europe	-0.005	-0.009	0.011	-0.007	0.001	0.005	-0.005	-0.005	-0.204	0.020	-0.140
	English Speaking	-0.001	-0.001	0.004	0.001	0.002	-0.004	0.003	-0.001	-0.021	0.076	-0.021
	Catholic Europe	-0.021	-0.008	0.002	0.0001	-0.003	-0.001	-0.004	0.003	-0.210	0.056	-0.234
	Confucian	-0.022	-0.017	0.025	0.001	-0.008	0.039	-0.008	-0.014	-0.458	0.006	-0.463
	West&South Asia	0.000	-0.002	0.003	0.001	0.001	0.007	-0.006	-0.001	-0.136	0.112	-0.184
	Latin America	0.002	-0.001	0.003	0.002	0.002	0.003	-0.001	0.003	-0.207	0.191	-0.242
	African-Islamic	0.000	-0.010	0.000	-0.001	-0.004	0.010	-0.008	-0.006	-0.264	0.206	-0.399
	Orthodox Europe	-0.001	-0.007	0.011	0.000	0.000	0.009	-0.005	-0.002	-0.190	0.066	-0.227
Gemini	Protestant Europe	-0.007	0.002	0.007	-0.004	0.000	0.009	-0.001	0.001	-0.142	0.068	-0.224
	English Speaking	-0.002	0.001	-0.001	-0.002	0.000	0.003	0.004	0.000	0.055	0.025	-0.003
	Catholic Europe	-0.009	-0.003	0.010	0.000	-0.002	0.018	-0.005	-0.003	-0.154	0.187	-0.283
	Confucian	-0.034	-0.006	0.001	-0.001	-0.006	0.030	-0.004	-0.001	-0.361	0.004	-0.421
	West&South Asia	-0.008	-0.003	0.003	-0.003	0.000	0.009	-0.003	-0.004	-0.102	0.130	-0.178
	Latin America	0.000	-0.002	0.005	-0.002	-0.001	0.006	0.001	-0.001	-0.021	0.168	-0.232
	African-Islamic	-0.007	-0.004	0.001	-0.004	-0.003	0.008	-0.005	0.000	-0.199	0.207	-0.348
	Orthodox Europe	-0.005	-0.002	0.010	0.000	0.001	0.012	-0.003	-0.002	-0.085	0.063	-0.202

Table 13: **Aggregate Shifts For Age Attributes added in Intersections With Other Attributes** across 4 models. The green text represents those values with significant (p<0.05) highest positive intensity and red text represents those values with the significant (p<0.05) highest negative intensity across each model's prediction per emotion.

Model	Attribute	Emotion								EPITOME		
		anger	anticipation	disgust	fear	joy	sadness	surprise	trust	ER	IP	EX
Llama-3-70B	0-17	0.015	0.012	-0.005	0.013	0.003	0.005	0.015	-0.001	0.160	0.055	0.096
	18-24	0.001	0.004	0.000	0.000	0.000	-0.005	0.003	-0.002	0.051	0.015	0.071
	25-34	-0.001	-0.001	0.001	-0.004	0.001	-0.002	-0.001	0.000	0.02	0.002	0.069
	35-44	-0.004	-0.006	0.003	-0.010	-0.001	-0.002	-0.005	0.002	0.017	-0.0011	0.022
	45-54	-0.009	-0.009	0.004	-0.010	-0.001	-0.001	-0.006	0.003	0.035	0.0035	-0.005
	55+	-0.019	-0.021	0.007	-0.018	-0.003	0.003	-0.010	-0.001	0.121	0.077	0.015
GPT-4o	0-17	-0.006	0.002	-0.004	0.009	0.003	-0.002	0.007	0.000	0.046	-0.002	0.023
	18-24	0.003	-0.002	-0.002	-0.002	0.005	-0.009	0.006	-0.003	0.018	0.005	0.118
	25-34	0.000	-0.001	0.002	-0.002	0.002	-0.002	0.004	-0.002	0.019	0.005	0.119
	35-44	-0.001	-0.007	0.003	-0.005	0.000	-0.001	0.000	-0.002	0.018	0.005	0.119
	45-54	-0.004	-0.011	0.002	-0.006	0.001	0.003	-0.003	-0.002	0.018	0.005	0.119
	55+	-0.007	-0.018	0.000	-0.007	0.002	0.007	-0.005	-0.003	0.025	-0.002	-0.027
DeepSeek	0-17	0.001	0.019	-0.008	0.022	-0.003	-0.009	0.015	0.006	0.203	0.039	0.055
	18-24	0.003	0.011	-0.003	0.003	0.003	-0.015	0.010	0.004	0.106	0.064	0.102
	25-34	0.002	0.004	0.001	-0.001	0.004	-0.008	0.004	0.002	0.058	0.023	0.124
	35-44	0.001	-0.002	0.003	-0.003	0.005	-0.003	0.003	0.001	-0.005	0.024	0.075
	45-54	0.000	-0.005	0.003	-0.006	0.006	-0.001	0.001	0.000	-0.005	0.013	0.082
	55+	-0.007	-0.012	0.001	-0.004	0.004	0.007	-0.001	-0.003	-0.001	0.0188	0.015
Gemini	0-17	-0.007	0.017	-0.017	0.025	0.001	-0.001	0.017	-0.001	0.163	-0.069	0.077
	18-24	0.0000	0.009	-0.005	0.003	0.001	-0.003	0.007	-0.003	0.125	-0.058	0.072
	25-34	0.008	0.000	0.003	0.003	0.002	-0.001	0.001	-0.002	0.118	-0.066	0.114
	35-44	0.005	0.000	0.000	-0.001	-0.001	-0.001	0.001	-0.002	0.111	-0.063	0.097
	45-54	0.001	-0.005	0.002	-0.002	0.000	0.004	0.000	-0.003	0.083	-0.053	0.078
	55+	-0.003	-0.011	0.002	-0.002	-0.001	0.011	-0.001	-0.005	0.080	-0.019	0.031

Table 14: **Aggregate Shifts For Cultural Attributes added in Intersections With Other Attributes** across 4 models. The green text represents those values with significant ($p < 0.05$) highest positive intensity and red text represents those values with the significant ($p < 0.05$) highest negative intensity across each model's prediction per emotion.

Model	Attribute	Emotion								EPITOME		
		anger	anticipation	disgust	fear	joy	sadness	surprise	trust	ER	IP	EX
Llama-3-70B	male	0.003	0.001	-0.005	-0.014	0.001	-0.013	-0.002	0.001	0.038	-0.017	-0.004
	female	-0.006	0.002	-0.001	0.010	-0.001	0.005	0.000	0.001	0.148	-0.006	0.091
	non-binary	-0.005	0.011	-0.006	0.003	0.001	-0.004	0.001	-0.002	0.101	0.074	0.091
	gender-queer	-0.000	0.013	-0.005	0.005	0.003	-0.010	0.001	-0.005	0.101	0.074	0.091
GPT-4o	male	0.004	-0.001	-0.001	-0.003	-0.002	-0.005	0.001	-0.001	0.007	-0.001	0.045
	female	0.005	-0.003	0.002	0.004	0.000	-0.001	-0.004	-0.001	0.035	-0.001	0.040
	non-binary	0.008	0.005	0.001	0.003	-0.004	-0.005	-0.001	-0.001	0.021	-0.008	-0.032
	gender-queer	0.014	0.008	0.000	0.005	-0.004	-0.009	-0.002	-0.001	0.027	-0.005	-0.018
Deep Seek	male	0.005	0.000	-0.005	-0.009	0.000	-0.013	-0.003	-0.002	0.011	0.010	0.018
	female	0.000	0.002	0.000	0.013	-0.004	0.000	0.000	0.001	0.067	0.095	0.006
	non-binary	0.004	0.002	-0.002	0.003	-0.004	-0.003	0.002	0.002	0.051	0.0255	0.071
	gender-queer	0.009	0.005	-0.002	0.007	-0.005	-0.003	0.001	0.000	0.027	0.045	0.045
Gemini	male	0.005	0.001	0.003	-0.002	-0.001	-0.003	0.002	0.000	-0.031	0.022	-0.025
	female	0.004	0.004	0.004	0.008	-0.002	0.003	0.002	-0.001	0.056	0.055	0.004
	non-binary	0.006	0.006	-0.002	-0.002	-0.002	-0.003	0.003	0.001	-0.011	0.055	-0.034
	gender-queer	0.007	0.011	-0.003	-0.002	-0.002	-0.006	0.004	0.002	-0.031	0.080	-0.016

Table 15: **Examples of Personas** Recalled in the Base State

Category	Example
Profession-Related	A nursing student accused of reporting someone for cheating
	A student
	A teacher or educator
	An experienced scuba diver or thrill-seeker
Behavioural	A concerned individual
	A skeptical individual
	A person from a stable background who values nostalgia
	A concerned and empathetic individual
Topic Related	A person who is frustrated with their neighbor's behavior
	A person who has recently completed their B.Sc degree
	A student seeking a recommendation letter
	A victim of rumours and gossip

Table 16: **Topic to Attribute Ratio** is calculated to assess whether the model's response for the attribute is skewed towards the topic or the attribute's characteristic. The values highlighted in blue represent those attributes where the model is likely to generate a response focusing on its characteristic.

Persona	Topic to Attribute Ration			
	GPT-4o-Mini	Llama-70B	Gemini 2.0	DeepSeek
0-17	0.708	0.778	0.816	0.886
18-24	0.700	0.714	0.778	0.839
25-34	0.636	0.688	0.749	0.806
35-44	0.640	0.685	0.751	0.803
45-54	0.657	0.694	0.747	0.812
55+	0.693	0.726	0.762	0.839
male	0.603	0.690	0.751	0.804
female	0.656	0.721	0.750	0.820
non-binary	0.717	0.724	0.754	0.867
gender-queer	0.786	0.819	0.823	1.012
Protestant Europe	0.628	0.834	0.838	0.942
English Speaking	0.566	0.729	0.803	0.820
Catholic Europe	0.691	0.891	0.910	0.999
Confucian	1.021	1.589	1.137	1.227
West and South Asia	0.682	0.797	0.866	0.934
Latin America	0.715	0.898	0.946	1.021
African-Islamic	0.779	1.01	0.961	1.093
Orthodox Europe	0.678	0.845	0.873	0.945

$$TAV = \frac{\text{Variance of the attribute's embeddings from the base culture}}{\text{Variance of the attribute's embeddings from the attribute's mean}} \quad (1)$$

A ratio > 1 indicates that the response embeddings for the given attribute are centered around the attribute's characteristics and stereotypes, while a ratio < 1 reflects that the response is more likely to be topic dependent. We show the topic to attribute variance in Table 16.

J Real World Gallup Scores

We collect the emotion scores from the Emotion World Report (Gallup-Analytics), which presents a comprehensive study across 142 nations. To process this dataset into our cultures, we prepare a mapping according to the Inglehart–Welzel Cultural Map from the countries to culture. We then decompose the emotion words studied in the Gallup poll into Plutchik's 8 Basic emotions (Plutchik, 1980) which are used by NRC and perform an aggregated weighted emotion score for each culture as follows in Table 17.

Table 17: **Real World Affective Scores** for emotion categories across cultures according to the Gallup World Poll

Culture	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Protestant Europe	0.008	0.103	0.0	0.048	0.182	0.056	0.0	0.113
English Speaking	0.013	0.104	0.0	0.060	0.180	0.069	0.0	0.112
Catholic Europe	0.011	0.100	0.0	0.054	0.170	0.062	0.0	0.107
Confucian	0.011	0.090	0.0	0.043	0.166	0.043	0.0	0.097
West and South Asia	0.014	0.102	0.0	0.046	0.182	0.060	0.0	0.112
Latin America	0.013	0.111	0.0	0.068	0.190	0.082	0.0	0.118
African-Islamic	0.023	0.093	0.0	0.068	0.154	0.082	0.0	0.100
Orthodox Europe	0.016	0.094	0.0	0.054	0.151	0.068	0.0	0.104

As seen in Table 17, the intensities for disgust and surprise are 0 across all cultures, and this is due to the absence of these intensities in the original dataset (Gallup-Analytics)