A Notations and Definitions

In our analysis, the dependent variable is the cosine similarity between an LLM's and a human's response vector — either 36-dimensional (for MFQ-2) or 19-dimensional (for WVS). An observation is defined as one cosine similarity under a specific combination of the model, prompt language, persona cue, and human participant, denoted as (m, l, p, h), where:

- $m \in M$ identifies the LLM, with $O_m \in \{+1 : \text{Chinese}, -1 : \text{American}\}$ indicating its origin;
- $l \in L$ is the prompt language, with $L_{\text{prmt}} \in \{+1 : \text{Mandarin}, -1 : \text{English}\}$ indicating its sum-coded form;
- $p \in P$ is the persona cue, mapped to the sum-coded variable $P_m \in \{+1 : \text{Chinese}, 0 : \text{None}, -1 : \text{American}\}$;
- $h \in H$ is a human participant, with $N_h \in \{+1 : \text{Chinese}, -1 : \text{American}\}$ indicating their nationality;
- $d \in D = \{\text{MFQ2}, \text{WVS}\}$ is the employed survey dataset (including survey statements and human participants' ratings). $|d| = 36$ for MFQ2 and $|d| = 19$ for WVS.

Thus, the independent variables (condition variables) used in the analysis are $C = \{O_m, P_m, N_h, L_{\text{prmt}}\}$, and the total number of unique observations is $|\text{obs}| = |M| \times |L| \times |P| \times |D|$. The observation data (X, y) consists of a matrix $X \in \mathbb{R}^{(K+1)\times|\text{obs}|}$, where each column encodes the values of the $K + 1$ condition variables (including an intercept term) for one observation, and a vector $y \in \mathbb{R}^{|\text{obs}|}$, containing the cosine similarity values between each LLM and human participant under the corresponding conditions in X.

Regression estimator. $f_{\text{reg}} : \{X, y\} \to \widehat{\beta}$ is a function taking the observation data (X, y) and returning the regression coefficients $\widehat{\beta}$. The significance-testing function, $f_{\text{sig}} : \{\widehat{\beta}, X, y\} \to p^i \in P$, takes the observation data as well as the regression coefficients and returns p-values, p^i, corresponding to $\widehat{\beta}_i \in \widehat{\beta}$.

Human responses. For each $h \in H$ and $d \in \{\text{MFQ2}, \text{WVS}\}$, let $R_h = \{r_{h,1}, \ldots, r_{h,|d|}\}$ where $r_{h,i} \in \{1, \ldots, 10\}$ is the raw Likert rating on item i of d.

LLM responses. For each (m, l, p) and iteration t, let $R^t_{m,l,p} \in \mathbb{R}^{|d|}$ be the response vector of the model m under the conditions of l, p. For each condition vector, the LLM is prompted $T = 20$ times to average out the stochasticity of LLM's response sampling. Accordingly the LLMs' averaged moral vector is represented in Equation 1:

$$\bar{R}_{m,l,p} = \frac{1}{T}\sum_{t=1}^{T} R^t_{m,l,p} \tag{1}$$

A.1 Normalization, Similarity, and Regression Models

Z-score normalization. When responding to the MFQ-2 items, LLMs gave higher ratings (*M* = 4.00) than human participants (*M* = 3.55). This difference is significant in a t-test: $t(37{,}582) = 10.3$, $p < 2e\text{-}16$. Among these human participants, Chinese participants gave higher ratings (*M* = 3.59) than American participants (*M* = 3.51): $t(37{,}150) = 6.5$, $p = 8e\text{-}11$.

The MFQ-2 is designed to identify the relative importance of moral dimensions (e.g., whether someone values equality more than loyalty), and these high/low biases obscure relative importance. However, a tendency to give higher or lower ratings across all dimensions can cause spurious differences in cosine similarity, obscuring relative importance.

We therefore z-scored ratings within each participant and within each condition for each LLM, such that each participant and each LLM has a mean normalized rating of 0, and positive values indicate a tendency to give higher ratings to questions in that dimension. Accordingly, for each human participant h, the Z-score normalized response vector is:

$$\mu_h = \frac{1}{|d|}\sum_{i=1}^{|d|} r_{h,i},$$
$$\sigma_h = \sqrt{\frac{1}{|d|}\sum_{i=1}^{|d|}(r_{h,i}-\mu_h)^2},$$
$$z_{h,i} = \frac{r_{h,i}-\mu_h}{\sigma_h}, \quad i = 1,\ldots,|d|,$$
$$z_h := \left(z_{h,1},\, z_{h,2},\, \ldots,\, z_{h,|d|}\right).$$

Similarly, for each LLM:

$$\mu_{m,l,p} = \frac{1}{|d|}\sum_{i=1}^{|d|} \bar{R}_{m,l,p,i},$$
$$\sigma_{m,l,p} = \sqrt{\frac{1}{|d|}\sum_{i=1}^{|d|}\left(\bar{R}_{m,l,p,i}-\mu_{m,l,p}\right)^2},$$
$$z_{m,l,p,i} = \frac{\bar{R}_{m,l,p,i}-\mu_{m,l,p}}{\sigma_{m,l,p}}, \quad i = 1,\ldots,|d|,$$
$$z_{m,l,p} := \left(z_{m,l,p,1},\, z_{m,l,p,2},\, \ldots,\, z_{m,l,p,|d|}\right).$$

Similarity between two z-vectors. Given any two z-vectors $z_{m,l,p}$ and z_h, we measure similarity using cosine similarity according to Equation 2:

$$\mathrm{Cos}\big(z^1, z^2\big) = \frac{\sum_{i=1}^{|d|} z_{m,l,p,i}\, z_{h,i}}{\sqrt{\sum_{i=1}^{|d|} z_{m,l,p,i}^2}\,\sqrt{\sum_{i=1}^{|d|} z_{h,i}^2}} \tag{2}$$

A.2 Regression Analysis

We perform linear least squares regression on the cosine similarity scores of each human and LLM z-vector to see how much each condition variable (such as LLM origin) and their interactions (such as LLM country of origin × participant nationality or congruence of nationality with LLM origin, system prompt persona, or prompt language) affect similarity to human participants.

The regression equation is:

$$y(h,m,l,p) = \mathrm{Cos}(\theta_{h,m,l,p}) = \beta_0 + \sum_{k=1}^{K}\beta_k\, x_k + \varepsilon_{h,m,l,p} \tag{3}$$

The design matrices X and y are:

$$X = \begin{pmatrix} 1 & x_1^1 & \ldots & x_K^1 \\ 1 & x_1^2 & \ldots & x_K^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{|\mathrm{obs}|} & \ldots & x_K^{|\mathrm{obs}|} \end{pmatrix}, \quad y = \begin{pmatrix} \mathrm{Cos}^1_{(\theta_{h,m,l,p})} \\ \mathrm{Cos}^2_{(\theta_{h,m,l,p})} \\ \vdots \\ \mathrm{Cos}^{|\mathrm{obs}|}_{(\theta_{h,m,l,p})} \end{pmatrix} \tag{4}$$

According to Equation 2, $y = X\beta + \epsilon$. Thus, the estimated regression coefficients $\widehat{\beta}$ are:

$$\widehat{\beta} = (X^T X)^{-1} X^T y \tag{5}$$

B Investigation of LLM response robustness

C RQ2 algorithm

: Investigating Congruence Effects (RQ2)

Require: human and LLM z-vectors, $\{z_h\}, \{z_{m,\ell,p}\}$.

Ensure: Regression coefficients $\hat{\beta}$ and p-values.

Form observation set $\mathcal{O}$ of tuples $(z_h,\, z_{m,\ell,p},\, C_L,\, C_P,\, C_O)$ where:
$C_L = \begin{cases} +1 & \text{if language } \ell \text{ matches } N_h, \\ -1 & \text{otherwise} \end{cases}$
$C_P = \begin{cases} +1 & \text{if persona } p \text{ matches } N_h, \\ -1 & \text{otherwise} \end{cases}$
$C_O = \begin{cases} +1 & \text{if LLM origin matches } O_m, \\ -1 & \text{otherwise} \end{cases}$
Construct $X \in \mathbb{R}^{|obs|\times 8}$ and $y \in \mathbb{R}^{|obs|}$
for $i = 1$ to $|obs|$ **do**
    Extract (C_L^i, C_P^i, C_O^i) from the i-th observation
    $X_i \leftarrow \big[1,\, C_L^i,\, C_P^i,\, C_O^i,\, C_L^iC_P^i,\, C_L^iC_O^i,\, C_P^iC_O^i,\, C_L^iC_P^iC_O^i\big],$
    $y_i \leftarrow \mathrm{Cos}\big(z_h^{(i)},\, z_{m,\ell,p}^{(i)}\big)$
end for
Compute the regression coefficients: $\hat{\beta} \leftarrow f_{reg}(X,y) \in \mathbb{R}^8$
For $j\{0,\ldots,7\}$, compute the p_values, p_j corresponding to each β_j: $P \leftarrow f_{sig}(\hat{\beta}, X, y) \in \mathbb{R}^8$
return $\hat{\beta}, P$

Interpretation of coefficients:

(a) If $p_1 < 0.05$ and $\hat{\beta}_1 > 0$, language congruence increases similarity.

(b) If $p_2 < 0.05$ and $\hat{\beta}_2 > 0$, persona congruence increases similarity.

(c) If $p_3 < 0.05$ and $\hat{\beta}_3 > 0$, origin congruence increases similarity.

(d) If $p_4 < 0.05$ and $\hat{\beta}_4 > 0$, synergy of language×persona adds $\hat{\beta}_4$[1].

(e) If $p_5 < 0.05$ and $\hat{\beta}_5 > 0$, synergy of language×origin adds $\hat{\beta}_5$.

(f) If $p_6 < 0.05$ and $\hat{\beta}_6 > 0$, synergy of persona×origin adds $\hat{\beta}_6$.

(g) If $p_7 < 0.05$ and $\hat{\beta}_7 > 0$, synergy of all three adds $\hat{\beta}_7$.

D RQ3 Algorithm

: Assessing Dimension Effects on Similarity (RQ3)

Require: $\{z_h\}$: human z-vectors, $\{z_{m,\ell,p}\}$: LLM z-vectors

Require: Dimensions $D = \{d_1, \ldots, d_K\}$ (e.g., Care, Equality for MFQ2 and suicide, abortion for WVS, coded as one-hot removal indicators, where K=6 for MFQ2 and K=19 for WVS.)

Require: Participant nationality code $N_h = 1$ (Chinese) and $N_h = -1$ (American) for each observation.

Ensure: Regression coefficients $\hat{\beta}$ and corresponding p-values, P.

For each observation $i \in obs$, compute $s_i = \mathrm{Cos}\big(z_h^{(i)}, z_{m,\ell,p}^{(i)}\big)$
for each dimension $d_k \in D$ **do**
  for each observation i **do**
    Remove items in d_k from $z_h^{(i)}$ and $z_{m,\ell,p}^{(i)}$
    Compute $\bar{s}_{i,k} = \mathrm{Cos}\big(z_h^{(i)} \setminus d_k,\ z_{m,\ell,p}^{(i)} \setminus d_k\big)$
    Set $\Delta_{i,k} = s_i - \bar{s}_{i,k}$
  end for
end for
for $j = 1$ to $|obs|$ **do**
  Let $k(j)$ be the dimension index for observation j.
  Set response $y_j = \Delta_{j,k(j)}$
  Construct row vector:

$$x_j = \big[\,1,\ \mathbb{I}(k(j)=1), \ldots, \mathbb{I}(k(j)=K),\ N_h^j\,\mathbb{I}(k(j)=1), \ldots, N_h^j\,\mathbb{I}(k(j)=K)\big]$$

end for
Compute the regression coefficients: $\hat{\beta} \leftarrow f_{reg}(X, y) \in \mathbb{R}^{2K+1}$
for $j \in \{0, \ldots, 2K+1\}$ compute the p_values corresponding to each β_j: $P \leftarrow f_{sig}(\hat{\beta}, X, y) \in \mathbb{R}^{2K+1}$

Interpretation:
for $k = 1$ to $2K+1$ do: **do**
  $\hat{\beta}_{1+k}$: effect of removing d_k (main effect). If $p_{1+k} < 0.05$ and $\hat{\beta}_{1+k} > 0$, excluding d_k increases similarity of LLM to humans overall.
  $\hat{\beta}_{1+K+k}$: interaction with nationality. If $p_{1+K+k} < 0.05$ and $\hat{\beta}_{1+K+k} > 0$, then removing d_k benefits similarity to Chinese participants more than to American participants.
end for