World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3102

PROBING THE COGNITIVE APPRAISAL STRUCTURE OF EMOTION REPRESENTATIONS IN LARGE LANGUAGE MODELS (LLMS): A FULLY AUTOMATED GEOMETRIC ANALYSIS BASED ON CHINESE–ENGLISH BILINGUAL CORPORA

HuaYing Liu[1], LiJie Luo[2*]
[1]*Independent Researcher, Guangzhou 510000, Guangdong, China.*
[2]*Guangxi Liyang Artificial Intelligence Application Software Co., Ltd., Nanning 530000, Guangxi, China.*
[*]*Corresponding Author: LiJie Luo*

Abstract: The emotional capabilities of large language models (LLMs) have been extensively validated, yet the organizational principles and cognitive regularities underlying their internal emotion representations remain poorly understood. This study introduces Smith and Ellsworth's (1985) cognitive appraisal theory and designs an automated experimental pipeline in which five mainstream LLMs rate Chinese and English emotional texts along six appraisal dimensions. Combining representational similarity analysis, principal component analysis, and unsupervised clustering, we systematically probed the geometric structure of the LLMs' emotion space. The results show that the appraisal structures of all LLMs are significantly aligned with human templates; however, under the English condition, the alignment did not surpass a purely semantic baseline, whereas under the Chinese condition the LLMs' appraisals significantly outperform the semantic baseline. The LLMs' appraisal space exhibits a systematic "responsibility shift," in which the responsibility dimension is disproportionately amplified on higher-order principal axes. Chinese alignment is significantly higher than English alignment for all models, and clustering structures were driven more by linguistic properties than by model architecture differences. These findings reveal that LLMs form emotion-cognitive structures that exhibit both commonalities and language-specific particularities, providing a psychology-theory-driven quantitative framework for explainable artificial intelligence (AI).
Keywords: Large language models; Emotion representations; Cognitive appraisal theory; Representational similarity analysis; AI cognition

1 INTRODUCTION

1.1 From Emotion Recognition to Cognitive Structure: An Unanswered Question

In recent years, large language models (LLMs) have demonstrated remarkable abilities in emotion analysis tasks, including emotion recognition [1], empathic response generation [2], and affective dialogue management [3]. These advances have spurred the exploration of LLM applications in affect-intensive settings such as psychological counseling and customer service. However, most existing studies have focused on the emotion labels or basic dimensions (e.g., valence-arousal-dominance, VAD) output by LLMs [4], failing to answer a deeper question: According to what organizational principles are the internal emotion representations of LLMs constructed? Do LLMs, like humans, understand and generate emotions through the underlying mechanism of "cognitive appraisal"?
Answering this question is not only relevant to the theoretical understanding of AI "cognition" but also has practical implications for building explainable and calibratable emotional AI—if we know on which appraisal dimensions LLMs diverge from human patterns, we can make targeted adjustments when designing interaction scenarios.

1.2 Cognitive Appraisal Theory: An Ideal Psychological Lens

Cognitive appraisal theory posits that emotions are not direct reflexes to external stimuli but rather the product of a series of cognitive appraisals that individuals make about events [5]. Through systematic experimental research, Smith and Ellsworth (1985) proposed a six-dimensional appraisal structure of emotional experience: pleasantness, anticipated effort, certainty, attention, responsibility, and situational control. These six dimensions constitute a multidimensional space in which different emotional states (e.g., anger, sadness, fear) occupy distinct positions.
This theoretical framework provides an ideal tool for probing the emotional organizing principles of LLMs. Compared with discrete emotion labels or VAD dimensions, cognitive appraisal theory offers the following advantages: First, it penetrates to the underlying cognitive structure of emotions—appraisal dimensions are the "causes" of emotions rather than their "descriptions"; Second, it provides a quantifiable multidimensional space suitable for geometric analysis; Third, it has been extensively validated in cross-cultural psychology [6], providing a theoretical basis for cross-linguistic analysis.

1.3 Research Aims and Core Questions

Using cognitive appraisal theory as a lens, this study systematically probes the organizational principles of LLMs' emotion representations and poses three core research questions:
RQ1: To what extent did the appraisal structures of LLMs align with typical human appraisal templates, and does this alignment surpass the level of purely semantic models?
RQ2: Do the LLMs' appraisal spaces exhibit cross-model shared organizational dimensions that are independent of the human template? If so, in what ways are they specific?
RQ3: Do the LLMs' emotional appraisal structures show systematic differences between Chinese and English contexts? If so, in what direction and with what possible explanations?
To achieve these aims, we designed an automated experimental pipeline that uses API calls to drive five representative LLMs to produce six-dimensional appraisal ratings on Chinese-English bilingual corpora, and we combined representational similarity analysis (RSA) [7], principal component analysis (PCA), and unsupervised clustering to quantitatively characterize the geometric structure of LLMs' emotion spaces.

2 METHOD

2.1 Experimental Methods

Corpus Preparation. The English corpus was drawn from the GoEmotions dataset [8], which contains 58,000 Reddit comments annotated with 27 fine-grained emotion categories. From this dataset, we selected sentences corresponding to six basic emotions (Ekman's classification [9]): sadness, anger, fear, disgust, joy, and surprise. Selection was accomplished through scripted random balanced sampling to ensure equal sample sizes across categories, yielding a final set of 512 English sentences.
The Chinese corpus was drawn from the SMP2020 Weibo Emotion Classification dataset released for the SMP2020 technical evaluation [10]. This dataset was annotated by the Research Center for Social Computing and Information Retrieval at Harbin Institute of Technology, with original Weibo data provided by the Micro-Hotspot Big Data Research Institute. The annotation labels include happiness, sadness, anger, fear, surprise, and neutral (no emotion). Notably, the label system does not cover the "disgust" category; therefore, the Chinese subset used in this study includes only five basic emotions: sadness, anger, fear, joy, and surprise. Scripted random balanced sampling yielded 512 Chinese sentences, to which Valence Arousal Dominance (VAD) coordinates and emotion category labels were appended for subsequent analyses. Robustness checks (see Methods section 2.2, Results section 3.6) confirmed that the absence of the "disgust" category had no substantial impact on the core conclusions.
Collection of LLM Appraisal Profiles. Five representative models were selected: GPT-5.5, Claude Opus 4.7, Gemini 3.1 Pro, DeepSeek V4, and Grok-4.20. These models span the current mainstream LLMs, including both closed-source and open-source types.
A uniform prompt template was used for all models, asking each model to generate six-dimensional appraisal ratings for each sentence. For example, the English prompt was:
You are an emotion analysis expert. For the given statement, output exactly a JSON array of six integers (1–5) in this order: [pleasantness, anticipated_effort, certainty, attention, responsibility, situational_control]. No other text.
A parallel Chinese prompt was used. The key design feature was that no emotion labels or category information was provided in the prompts; only the six-dimensional ratings were requested. Appendix A presents an example API-calling script (using DeepSeek) for reviewer verification.
Semantic Baseline. The Zhipu Embedding-3 model was used to generate a general-purpose semantic vector for each sentence, serving as a non-appraisal-oriented pure semantic baseline for comparison with the LLMs' appraisal orientations.
Construction of the Human Reference Template. The emotion–appraisal dimension mapping method was employed. First, for the emotion_category label of each sentence in the corpus, the corresponding six-dimensional appraisal means were looked up. The appraisal means were taken from the updated meta-analytic norms of Yeo and Ong (2024) [11], which integrate appraisal data from 47 independent studies conducted between 1985 and 2023. For emotion categories not included in the norms (the missing "disgust" category in the Chinese subset), neutral default values were assigned. This produced an $N \times 6$ human appraisal template matrix.

2.2 Data Analysis Methods

The entire experiment was automatically driven by a set of Python scripts comprising four main steps: Step 1 (corpus preparation and balanced sampling), Step 2 (calling each LLM's Chat API for six-dimensional appraisal ratings), Step 3 (generating semantic baseline vectors), and Step 4 (core data analysis). Data analysis encompassed six phases, all completed within Step 4. All API-calling scripts and the complete analysis code have been open-sourced (see Appendix B for the repository URL). Appendix A provides an example of the core data collection script.
Data Cleaning and Missing Value Pre-checks. All LLM rating matrices were examined for missing values. Missing dimension ratings were imputed with the median of the corresponding model–language–dimension combination. Subsequently, a missing-value sensitivity analysis was performed: for model–language combinations with a missing

rate greater than 5%, the core correlation coefficients between LLM and human ratings were recomputed after removing missing sentences, and the results were compared with those based on the imputed data.

Construction of Representational Similarity Matrices (RSMs). Cosine similarities between sentences were computed separately for the LLM appraisal matrices, the human template matrix, and the semantic baseline vectors, generating three types of RSMs (LLM, human, and semantic). To fully characterize the information sources of each RSM, three reference matrices—sentence length difference, term frequency–inverse document frequency (TF-IDF) cosine similarity, and emotion label Jaccard similarity—were simultaneously computed as benchmarks for subsequent RSA.

Core Representational Similarity Analysis (RSA). Spearman rank correlations between each model's RSM and the human RSM were computed, yielding the correlation coefficient between LLM and human ratings, r_s(LLM, human). Correlations between each model's RSM and the semantic RSM gave r_s(LLM, Semantic). Mantel permutation tests (5,000 permutations, False Discovery Rate (FDR) correction) were used to test whether the LLM–human correlation was significantly greater than zero. Steiger's Z test was employed to directly compare LLM–human and semantic–human correlations [12]. Cross-model and cross-language analyses were conducted with Fisher's Z test.

Geometric Analysis. Pearson correlation matrices among the LLMs' appraisal dimensions were computed, and the RV coefficient was used to quantify the overall similarity of the dimension covariation patterns to those of the human template [13]. PCA was performed on the LLM appraisal matrices, extracting PC1 and PC2 to analyze loading shifts of key dimensions.

Unsupervised Clustering. Hierarchical clustering (Ward's method, with the optimal K determined by the silhouette coefficient) was conducted on the LLMs' RSMs, and the clustering results were compared with the true emotion labels using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). One-way ANOVA was used to analyze differences among clusters on each appraisal dimension, with effect size η^2 identifying the key dimensions driving cluster organization.

Robustness Checks. Two checks were performed for distinct purposes. First, the original norms of Smith and Ellsworth (1985) replaced the updated norms of Yeo and Ong (2024) to test the sensitivity of the core conclusions to norm selection [5,11]. Second, disgust sentences were removed from the English corpus to eliminate the structural asymmetry introduced by the absence of the disgust category in the Chinese corpus, ensuring fairness of cross-linguistic comparison.

3 RESULTS

3.1 Data Cleaning and Missing Value Analysis

The missing rates across all 10 model–language conditions were summarized in Table 1. The only condition with a notably high missing rate was DeepSeek V4 on the English corpus (6.64%). Sensitivity analysis showed that after removing the 34 missing sentences, the change in the core correlation coefficient between LLM and human ratings relative to the imputed data was merely –0.006 ($|\Delta r| < 0.02$), confirming that median imputation did not distort the core findings.

Table 1 Missing Rate Statistics Across Model–Language Conditions

Model	Language	Total sentences	Missing sentences	Missing rate (%)
Claude Opus 4.7	EN	512	3	0.59
	ZH	512	1	0.20
DeepSeek V4	EN	512	34	6.64
	ZH	512	0	0
Gemini 3.1 Pro	EN	512	2	0.39
	ZH	512	0	0
GPT-5.5	EN	512	1	0.20
	ZH	512	0	0
Grok-4.20	EN	512	3	0.59
	ZH	512	1	0.20

3.2 Associations Between Reference Matrices and RSMs

As a control, we examined the Spearman rank correlations of the three reference matrices with each RSM. Sentence length difference and TF-IDF cosine similarity were not significantly correlated with any RSM. The Jaccard similarity of emotion category labels was highly correlated with the human template RSM (see Table 2) but showed only weak positive correlations with the LLMs' RSMs, consistent with the expected natural association between text semantics and emotion categories.

Table 2 Spearman Correlations Between Emotion Label Jaccard Similarity and RSMs

RSM type	Reference matrix	Language	Spearman *r*	*p*
Human template	Sentence length	EN	–0.003	0.239
Human template	TF-IDF	EN	–0.007	0.010
Human template	Label Jaccard	EN	0.644	< .001
Claude Opus 4.7	Label Jaccard	EN	0.076	< .001
DeepSeek V4	Label Jaccard	EN	0.082	< .001
Gemini 3.1 Pro	Label Jaccard	EN	0.072	< .001
GPT-5.5	Label Jaccard	EN	0.082	< .001
Grok-4.20	Label Jaccard	EN	0.062	< .001
Human template	Label Jaccard	ZH	0.693	< .001
Claude Opus 4.7	Label Jaccard	ZH	0.217	< .001
DeepSeek V4	Label Jaccard	ZH	0.174	< .001
Gemini 3.1 Pro	Label Jaccard	ZH	0.189	< .001
GPT-5.5	Label Jaccard	ZH	0.231	< .001
Grok-4.20	Label Jaccard	ZH	0.225	< .001

3.3 Core RSA: Universal Alignment and Cross-linguistic Divergence

The core RSA results are shown in Table 3. In all 10 model–language conditions, the correlation coefficients between LLM and human ratings were significantly positive (Mantel permutation test, FDR-corrected $p < .001$), confirming that the LLMs' appraisal structures share common information with the human template.

Steiger's Z test revealed a clear cross-linguistic asymmetry. Under the Chinese condition, the LLM–human correlations were significantly higher than the semantic–human correlations for all five models (Z values ranging from 3.03 to 4.76, all $p < .01$), indicating that the LLMs' appraisal alignment surpassed similarity at the purely semantic level. Under the English condition, the differences between LLM–human and semantic–human correlations were not significant for any of the five models (Z values ranging from 0.23 to 0.68, all $p > .05$).

Table 3 Core RSA Results

Model	Language	r_s(LLM,Human)	*p*(LLM,Human)	r_s(LLM,Semantic)	*p*(LLM,Semantic)	Steiger's Z	Steiger's *p*
Claude	EN	0.12	< .001	0.08	< .001	0.68	0.499
	ZH	0.35	< .001	0.07	< .001	4.76	< .001
DeepSeek	EN	0.13	< .001	0.10	< .001	0.48	0.633
	ZH	0.27	< .001	0.08	< .001	3.03	0.003
Gemini	EN	0.08	< .001	0.06	0.001	0.26	0.794
	ZH	0.27	< .001	0.06	0.001	3.54	< .001
GPT	EN	0.13	< .001	0.12	< .001	0.23	0.819
	ZH	0.34	< .001	0.11	< .001	3.93	< .001
Grok	EN	0.08	< .001	0.05	0.001	0.46	0.646
	ZH	0.37	< .001	0.12	< .001	4.32	< .001

Note. r_s denotes the Spearman rank correlation coefficient. r_s(LLM,Human) is the correlation between LLM and human ratings; r_s(LLM,Semantic) is the correlation between LLM and the semantic baseline.

Cross-linguistic comparisons are presented in Table 4. Chinese alignment was significantly higher than English alignment for all models (Fisher's Z test, $p < .05$). This cross-model consistent phenomenon is highly robust (Figure 1).

Table 4 Cross-Language Comparison: English vs. Chinese Alignment by Model

Model	r_s(EN)	r_s(ZH)	Fisher's Z	*p*
Claude Opus 4.7	0.12	0.35	–3.87	< .001
DeepSeek V4	0.13	0.27	–2.27	0.023
Gemini 3.1 Pro	0.08	0.27	–3.25	0.001
GPT-5.5	0.13	0.34	–3.62	< .001
Grok-4.20	0.08	0.37	–4.93	< .001

Note. r_s is the Spearman rank correlation coefficient between LLM and human ratings.

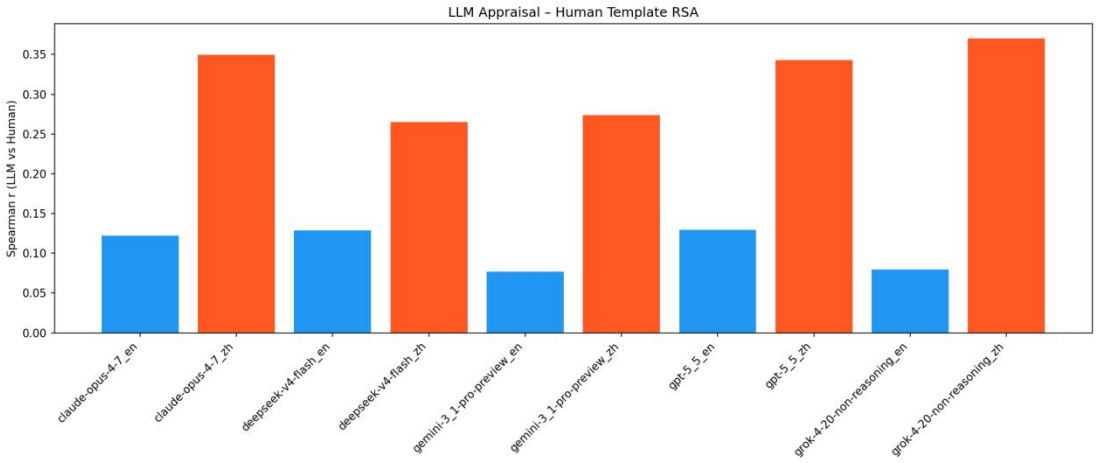

Figure 1 Cross-Language RSA Results

3.4 Geometric Structure of the Appraisal Space: A Shared Primary Axis and Shifted Secondary Axes

Dimension Covariation Network. Table 5 presents the RV coefficients for each model, quantifying the overall similarity of their dimension covariation patterns to the human template. The mean RV coefficient for the English models (M = 0.13, SD = 0.01) was significantly lower than that for the Chinese models (M = 0.35, SD = 0.03), independent-samples t-test: $t(8) = 16.63$, $p < .001$, indicating that the inter-dimension covariation patterns in the LLMs' English appraisal space deviate more substantially from the human template (Figure 2).

Table 5 RV Coefficients: Dimension Covariation Between LLM and Human Template

Model	English RV	Chinese RV
Claude Opus 4.7	0.13	0.36
DeepSeek V4	0.14	0.30
Gemini 3.1 Pro	0.13	0.34
GPT-5.5	0.14	0.39
Grok-4.20	0.11	0.34

Note. RV reflects the overall similarity between the LLM and human dimension correlation matrices.

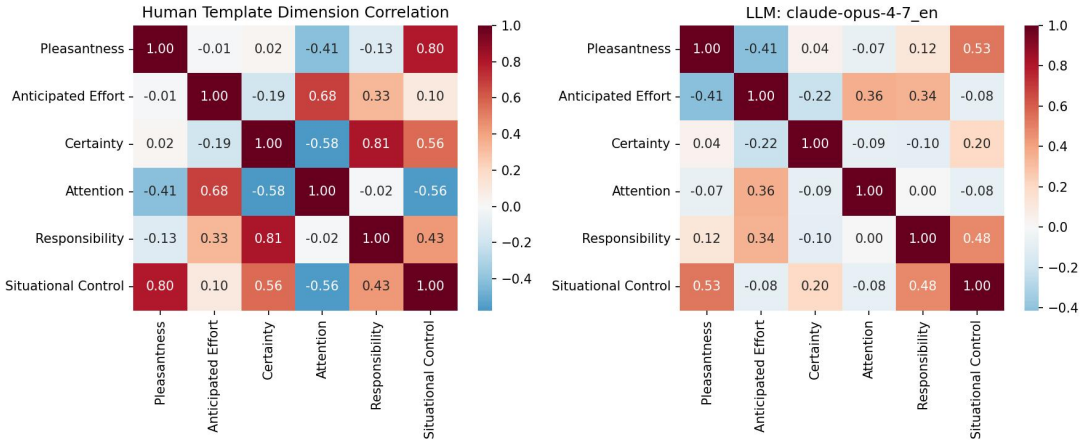

Figure 2 Dimension Correlation Heatmap: Human vs. LLMs

Principal Component Analysis (PCA). Table 6 shows the loadings of the key dimensions—pleasantness, responsibility, and certainty—on PC1 and PC2 for each model. PC1 was dominated by pleasantness for all models (loadings ranging from 0.64 to 0.81), forming a shared valence primary axis. The structure of PC2 displayed a systematic AI-specific pattern: except for Grok-4.20 under the English condition, PC2 was dominated by responsibility for all models in both languages (loadings ranging from 0.65 to 0.94), whereas in the typical human template the contribution of responsibility to PC2 is minimal. We conceptualize this phenomenon as the "responsibility shift," the most salient specificity in the LLMs' appraisal structure. Grok-4.20 under the English condition was an exception, with PC2 dominated by certainty (loading 0.74; its responsibility loading was only 0.06), exhibiting a unique organizational logic. However, under the

Chinese condition it still showed PC2 dominance by responsibility (loading 0.94), indicating that linguistic context exerts a stronger influence on Grok-4.20's appraisal structure (Figure 3).

Table 6 PCA Loadings: Key Dimensions on PC1 and PC2

Model	Language	PC1 dominant dimension	PC1 loading	PC2 dominant dimension	PC2 loading
Claude Opus 4.7	EN	Pleasantness	0.81	Responsibility	0.65
	ZH	Pleasantness	0.81	Responsibility	0.72
DeepSeek V4	EN	Pleasantness	0.69	Responsibility	0.77
	ZH	Pleasantness	0.64	Responsibility	0.79
Gemini 3.1 Pro	EN	Pleasantness	0.75	Responsibility	0.78
	ZH	Pleasantness	0.71	Responsibility	0.74
GPT-5.5	EN	Pleasantness	0.80	Responsibility	0.78
	ZH	Pleasantness	0.77	Responsibility	0.80
Grok-4.20	EN	Pleasantness	0.73	Certainty	0.74
	ZH	Pleasantness	0.72	Responsibility	0.94

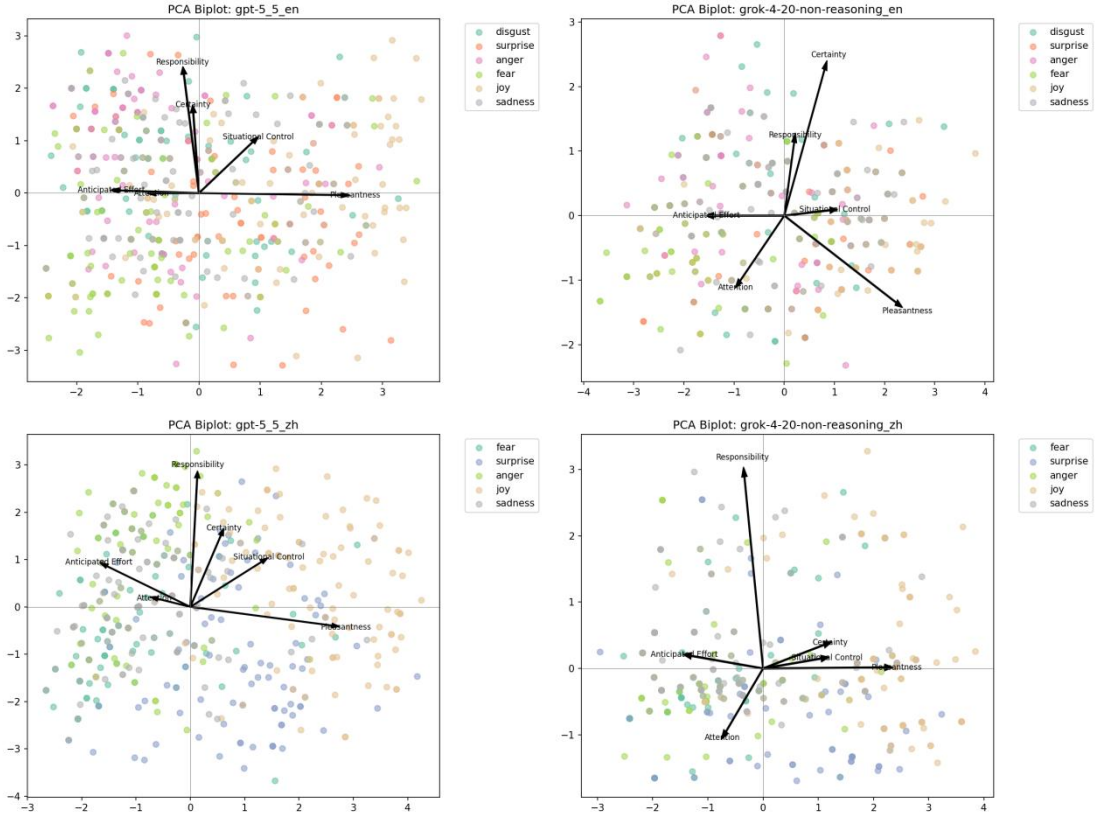

Figure 3 Example PCA Biplots

3.5 Unsupervised Clustering: AI Emotion Self-organization

Cluster Purity. Hierarchical clustering results showed that the cluster purity (NMI and ARI) was low for all models (Table 7), indicating that the LLMs' emotion clusters are not a simple reproduction of human emotion categories. The NMI values for the English models ranged from 0.08 (Claude Opus 4.7) to 0.09 (GPT-5.5 and Gemini 3.1 Pro), and for the Chinese models from 0.14 (Gemini 3.1 Pro) to 0.24 (GPT-5.5). Chinese cluster purity was slightly higher, but the overall level remained low.

Table 7 Cluster Purity Summary

Model	Language	Optimal K	NMI	ARI
Claude Opus 4.7	EN	3	0.08	0.05
	ZH	2	0.23	0.14
DeepSeek V4	EN	3	0.08	0.06
	ZH	3	0.17	0.14

Gemini 3.1 Pro	EN	3	0.09	0.06
	ZH	2	0.14	0.07
GPT-5.5	EN	4	0.09	0.07
	ZH	2	0.24	0.15
Grok-4.20	EN	5	0.08	0.05
	ZH	2	0.21	0.16

Cluster Driving Forces. ANOVA effect sizes (η^2) revealed the key appraisal dimensions driving clustering (Table 8). Pleasantness was the most stable cross-language, cross-model clustering driver (η^2 ranging from 0.52 to 0.77), ranking first or second in all 10 conditions. The effect size of certainty was extremely high in English (η^2 ranging from 0.53 to 0.70), but, except for DeepSeek ($\eta^2 = 0.69$), it dropped sharply in Chinese (η^2 ranging from 0.00 to 0.29), reflecting systematic differences in certainty expressions between Chinese and English corpora. The effect size of responsibility varied greatly among English models (Grok-4.20 $\eta^2 = 0.62$, GPT-5.5 $\eta^2 = 0.53$ vs. Claude Opus 4.7 $\eta^2 = 0.08$, DeepSeek V4 $\eta^2 = 0.04$), further corroborating the specificity and model-dependence of the responsibility dimension in LLMs' appraisal organization. In Chinese, Grok-4.20 showed an extremely high effect size for anticipated effort ($\eta^2 = 0.72$) with almost no contribution from responsibility ($\eta^2 = 0.01$), consistent with its PCA results (Figure 4).

Table 8 ANOVA Effect Sizes (η^2): Appraisal Dimension Contributions to Clustering

Model–Language	Pleasantness	Anticipated Effort	Certainty	Attention	Responsibility	Situational Control
Claude Opus 4.7 (EN)	0.65	0.14	0.62	0.04	0.08	0.28
Claude Opus 4.7 (ZH)	0.76	0.27	0.01	0.00	0.01	0.18
DeepSeek V4 (EN)	0.52	0.19	0.70	0.13	0.04	0.28
DeepSeek V4 (ZH)	0.68	0.23	0.69	0.11	0.04	0.33
Gemini 3.1 Pro (EN)	0.74	0.28	0.57	0.10	0.03	0.14
Gemini 3.1 Pro (ZH)	0.63	0.46	0.01	0.03	0.08	0.08
GPT-5.5 (EN)	0.76	0.27	0.53	0.15	0.53	0.23
GPT-5.5 (ZH)	0.77	0.34	0.00	0.12	0.01	0.21
Grok-4.20 (EN)	0.62	0.64	0.69	0.28	0.62	0.46
Grok-4.20 (ZH)	0.75	0.72	0.29	0.17	0.01	0.20

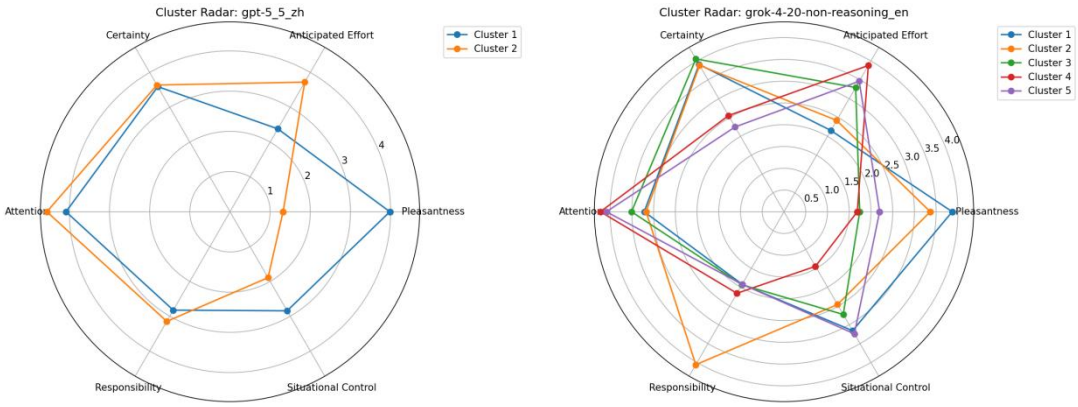

Figure 4 Example Cluster Radar Plots

Cross-model Cluster Consistency. The cross-model NMI matrix (Table 9) shows that the clustering structure sharing among different models within the same language is much higher than cross-language sharing. Within-language average NMI among English models = 0.24; among Chinese models = 0.38. Cross-language average NMI between any English and Chinese model pair = 0.03 (Figure 5). (Calculated from the full 10 × 10 cross-model NMI matrix; see Supplementary Table S1 in Appendix B.)

Table 9 Cross-Model Cluster Consistency

Model	Claude EN	Claude ZH	GPT EN	GPT ZH	Grok EN	Grok ZH
Claude Opus 4.7 EN	1	0.02	0.32	0.03	0.21	0.03
Claude Opus 4.7 ZH	0.02	1	0.03	0.48	0.02	0.44
GPT-5.5 EN	0.32	0.03	1	0.04	0.19	0.04
GPT-5.5 ZH	0.03	0.48	0.04	1	0.02	0.41

Grok-4.20 EN	0.21	0.02	0.19	0.02	1	0.02
Grok-4.20 ZH	0.03	0.44	0.04	0.41	0.02	1

Note. The full 10 × 10 matrix is provided in Supplementary Table S1, Appendix B.

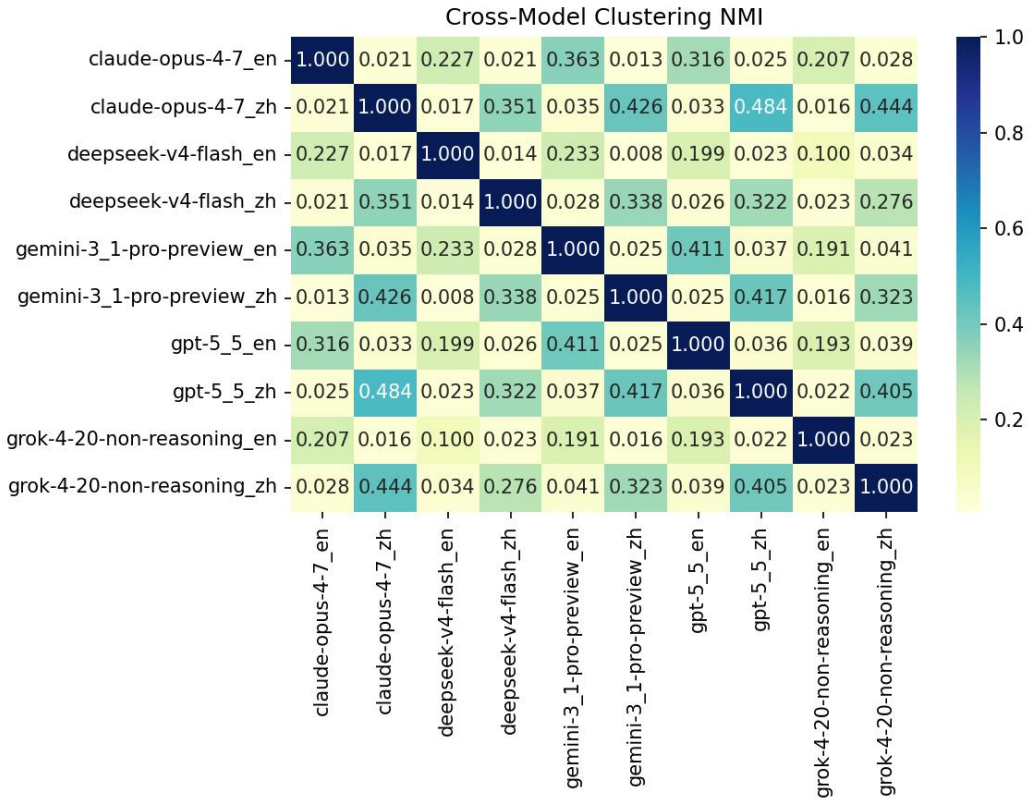

Figure 5 Cross-Model NMI Heatmap

3.6 Robustness Checks

Substituting Original Norms (testing sensitivity to norm choice). When the human reference template was replaced with the original norms of Smith and Ellsworth (1985), the recalculated model performance rankings were highly consistent with the main analysis results (Spearman r = 0.95, p < .001; Table 10), indicating that the core conclusions are insensitive to norm choice.

Table 10 Model Performance Under 1985 vs. 2024 Norms

Model	Language	1985 Norm r_s	2024 Norm r_s
Claude Opus 4.7	EN	0.13	0.12
	ZH	0.31	0.35
DeepSeek V4	EN	0.14	0.13
	ZH	0.24	0.27
Gemini 3.1 Pro	EN	0.08	0.08
	ZH	0.27	0.27
GPT-5.5	EN	0.13	0.13
	ZH	0.30	0.34
Grok-4.20	EN	0.07	0.08
	ZH	0.33	0.37

Note. Model rankings under the 1985 and 2024 norms are highly consistent (Spearman r = .95, p < .001).

Robustness Check for the Absence of "Disgust" in the Chinese Corpus (eliminating structural asymmetry). Because the SMP2020 Chinese dataset does not include the "disgust" category, an inherent asymmetry exists in the human–machine alignment: the English human template includes means based on disgust sentences, whereas the Chinese human template lacks corresponding data points. Although this omission only affects the placeholder value for the disgust category in the human template (neutral default values were assigned), its potential influence on the core conclusions still needed evaluation.

We therefore removed all disgust-labeled sentences from the English corpus (approximately 86 sentences, leaving 426 sentences), making the English corpus's category structure consistent with that of the Chinese corpus (both lacking disgust), and recalculated the core indicators. The results after removal are shown in Table 11. The changes in LLM–human correlation coefficients, |Δr|, were all less than 0.03 for every model, and the core conclusions underwent no substantive change, confirming that the absence of the disgust category in the Chinese corpus does not affect the main findings.

Table 11 Changes in Core Indicators After Removing Disgust Sentences

Model	*n* (after removal)	r_s(Original)	r_s(after removal)	Δ*r*
Claude Opus 4.7	426	0.12	0.11	–0.01
DeepSeek V4	426	0.13	0.13	0.00
Gemini 3.1 Pro	426	0.08	0.10	0.02
GPT-5.5	426	0.13	0.12	–0.01
Grok-4.20	426	0.08	0.10	0.02

Main Summary Table and Forest Plot. Table 12 summarizes the core indicators across all 10 conditions of the study, including RSA alignment (LLM–human correlation coefficients with their bootstrap 95% confidence intervals), cluster purity (NMI, ARI), and dimension covariation pattern similarity (RV coefficients), facilitating systematic cross-model and cross-linguistic comparisons for readers. Figure 6 presents these LLM–human correlation coefficients and their confidence intervals in a forest plot, visually depicting the distribution of effect sizes across languages and models.

Table 12 Main Summary Table: All Core Indicators

Model	Language	r_s(LLM,human)	Lower 95% CI	Upper 95% CI	NMI	ARI	RV
Claude Opus 4.7	EN	0.12	0.09	0.17	0.08	0.05	0.13
	ZH	0.35	0.30	0.40	0.23	0.14	0.36
DeepSeek V4	EN	0.13	0.09	0.18	0.08	0.06	0.14
	ZH	0.27	0.22	0.32	0.17	0.14	0.30
Gemini 3.1 Pro	EN	0.08	0.04	0.12	0.09	0.06	0.13
	ZH	0.27	0.23	0.33	0.14	0.07	0.34
GPT-5.5	EN	0.13	0.09	0.18	0.09	0.07	0.14
	ZH	0.34	0.29	0.40	0.24	0.15	0.39
Grok-4.20	EN	0.08	0.05	0.12	0.08	0.05	0.11
	ZH	0.37	0.32	0.43	0.21	0.16	0.34

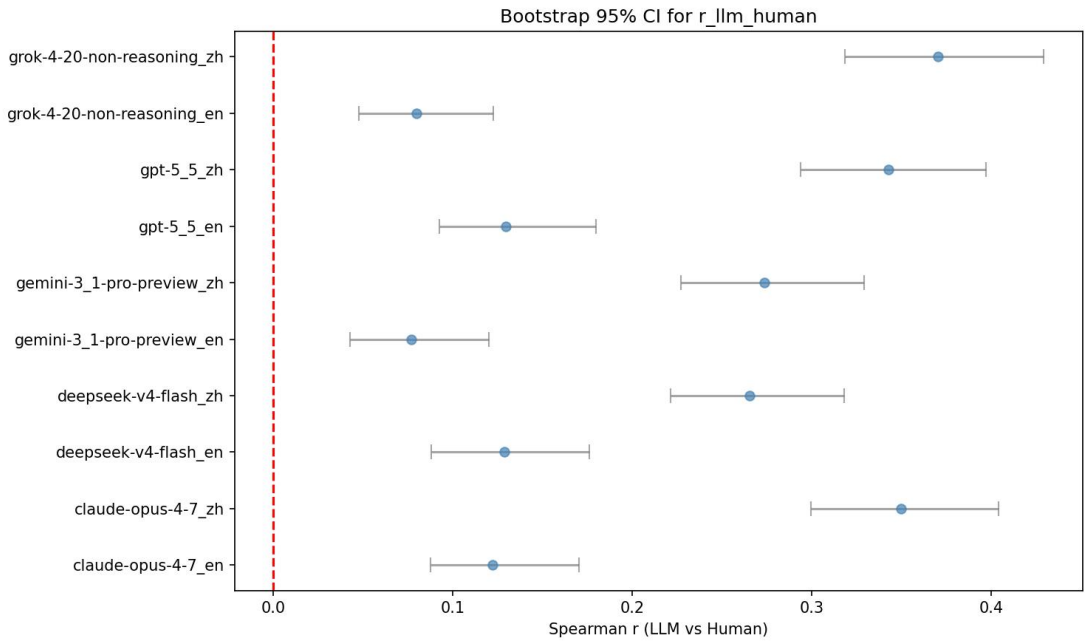

Figure 6 Forest Plot of Core Indicators

4 DISCUSSION

4.1 AI Emotion Geometry Through the Lens of Appraisal Theory

Through the lens of cognitive appraisal theory, this study revealed organizational principles of LLMs' emotion representations that exhibit both human-like commonalities and AI-specific particularities.

Regarding RQ1 (alignment), the appraisal structures of all LLMs were significantly aligned with the human template (Mantel permutation tests passed FDR correction for all model–language conditions, $p < .001$), and the valence dimension (pleasantness) dominated PC1 for all models in both languages (loadings ranging from 0.64 to 0.81), forming a cross-model shared "valence primary axis." This finding indicates that although LLMs do not "experience" emotions in the manner of humans, their six-dimensional appraisals of emotional sentences have internalized the most fundamental valence dimension of human emotional cognition in terms of structural organization. This is directionally consistent with prior studies showing LLMs' near-human performance on emotion labeling tasks [14], but our study further reveals that such human-like performance is relatively adequate on first-order tasks (label recognition) but diverges on second-order tasks (deep organization of appraisal structure) depending on language context. Under the English condition, the LLM–human correlation did not differ significantly from the semantic–human correlation (Steiger's Z test $p > .05$ for all five models), suggesting that in English, LLMs' appraisal alignment may partly stem from contributions of text semantic similarity. Under the Chinese condition, all models' LLM–human correlations were significantly higher than semantic–human correlations (Z values ranging from 3.03 to 4.76, all $p < .01$), indicating that the Chinese context can elicit a human-like appraisal structure that transcends the purely semantic level.

Regarding RQ2 (independence), alongside sharing the valence primary axis with humans, the LLMs' appraisal space also exhibited a systematic "responsibility shift"—the responsibility dimension was disproportionately amplified on PC2 (loadings ranging from 0.65 to 0.94), whereas in the typical human appraisal structure, PC2 is usually dominated by certainty or situational control, with minimal contribution from responsibility. We conceptualize this cross-model shared AI-specific pattern as the "responsibility shift" and regard it as the core organizational feature that distinguishes LLMs' emotion-cognitive structures from those of humans. This finding aligns with the tendency of LLMs to over-attribute internal causes in attribution tasks and may be rooted in stable linguistic patterns within the training corpora [15]: when texts describe negative events, the subject of the sentence is often directly or indirectly assigned responsibility.

However, the specific strength of the responsibility shift varies across models. Grok-4.20 was a notable exception under the English condition: its PC2 was dominated by certainty rather than responsibility (loading 0.74 vs. 0.06). We speculate that such variation may arise from differences in training data composition, architecture design, or post-training stages among models—for example, Grok's multi-agent debate architecture and large-scale reinforcement learning targeting low hallucination rates may have proactively biased its appraisal structure toward certainty-oriented outputs. Notably, under the Chinese condition, Grok-4.20's PC2 was still dominated by responsibility (loading 0.94), indicating that the modulatory role of language context can even override model-specific architecture. This precisely reinforces the study's overall argument: the appraisal structures of LLMs exhibit a cross-model shared shift direction, but the strength and even the type of shift are adjusted according to each model's uniqueness.

Integrating the above findings reveals a more general scenario: the emotion representations of LLMs are neither simple replicas of human patterns nor random noise, but rather a "quasi-cognitive structure"—one that involves both the projection of linguistic features and the existence of human-independent, algorithmically or data-driven specific organizational dimensions. To deepen the understanding of this structure, a key variable must be examined: whether language context systematically modulates the LLMs' appraisal patterns. This is precisely the question answered in the next section.

4.2 Cross-linguistic Differences and Their Cultural–Psychological Implications

In answering RQ3, this study uncovered a stable and cross-model-consistent effect: the appraisal alignment of all five LLMs was significantly higher under the Chinese condition than under the English condition (Fisher's Z test, all $p < .05$; see Table 4). This difference is not only statistically significant but also repeatedly emerges across multiple analyses, constituting one of the most robust findings of this experiment.

From the perspective of the overall similarity of dimension covariation patterns, the mean RV coefficient of the English models ($M = 0.13$) was significantly lower than that of the Chinese models ($M = 0.35$), independent-samples t-test: $t(8) = 16.63$, $p < .001$, with an extremely large effect size (Cohen's $d = 10.52$), indicating that the relational structure among dimensions in the English appraisal space deviates more substantially from the human template. In terms of cluster purity, the NMI values of the English models ranged from 0.08 to 0.09, while those of the Chinese models ranged from 0.14 to 0.24. Although the overall level remained low (confirming that LLMs do not simply reproduce human emotion categories), the clustering structure under the Chinese condition was clearly closer to the organizational manner of human emotion labels. Furthermore, cross-model cluster consistency analysis showed that within-language cluster sharing among models (average $NMI_{English} = 0.24$; $NMI_{Chinese} = 0.38$) was far greater than cross-language sharing (average NMI between any English and Chinese model pair = 0.03), directly demonstrating that language context—not model architecture—is the dominant force determining clustering structure.

Human cultural psychology provides a powerful explanatory framework for this cross-model-consistent cross-linguistic asymmetry. Zhang (2018) noted that language could serve as an effective cultural prime [16], influencing individuals' attributional styles, cognitive patterns, and personality expressions. His empirical research demonstrated that Chinese–English bilinguals tend toward environmental attributions when their Chinese identity is activated, and toward internal attributions when their American identity is activated. Our findings are highly consistent with this: emotional expressions in the Chinese corpus (Weibo) are generally more implicit, more emphatic of social relationships and responsibility ascriptions, structurally echoing the relatively higher weights of responsibility in certain emotion categories of the human template; emotional expressions in the English corpus (Reddit) are more individualistic and direct, which may lead the LLMs' appraisals to rely more on surface-level semantic features, thereby weakening alignment with the deep structure of human appraisals. Cultural frame switching theory further provides a cognitive–mechanistic explanation for this phenomenon [17]: although LLMs are not agents with cultural identities, the deeply embedded linguistic–cultural conventions within their training data—for instance, Chinese uses fewer absolute-judgment constructions, whereas English places greater emphasis on individual agency—may "activate" different inferential modes when the models appraise emotional sentences, making their behavior exhibit effects analogous to human cultural frame switching.

Existing research has found that when the prompting language is Chinese, mainstream LLMs exhibit a more interdependent social orientation and a more holistic cognitive style; when switched to English, they shift toward an independent orientation and an analytic style [18]. If the cultural tendencies of LLMs shift with language, then the phenomenon of "higher Chinese alignment" in this study acquires a possible explanation: under the Chinese context, the model's cultural default values are more congruent with the interdependent cultural framework on which the Chinese human appraisal template rests; under the English context, the model must operate within a cultural framework of lower compatibility, and alignment consequently decreases.

The cluster driving force analysis provides more microscopic evidence for the above language-level mechanisms. The certainty dimension was the second strongest clustering driver after pleasantness under the English condition (η^2 ranging from 0.53 to 0.70), but in most Chinese models its effect size dropped sharply to near zero (η^2 ranging from 0.00 to 0.29). This difference can be traced to systematic differences in certainty expressions between Chinese and English corpora: Chinese Weibo posts frequently use open-ended expressions such as "maybe," "possibly," or "perhaps," whereas English Reddit comments tend to make direct judgments with deterministic phrases such as "This is clearly" or "I am sure that." DeepSeek V4 still maintained a relatively high certainty effect size under the Chinese condition ($\eta^2 = 0.69$); this exception simultaneously suggests that even though language context is the strongest shared modulating variable at the macro level, individual models' sensitivity to specific linguistic features still varies. Such "language effect × model specificity" interactions are a direction worthy of deeper exploration in future research.

In sum, the core message of the cross-linguistic differences is that language context—possibly through the deep entanglement of language and cultural conventions in the training data—is the strongest variable modulating the LLMs' appraisal structures, and its influence even exceeds differences stemming from model architectures.

4.3 Methodological Contribution: A Modular, Collaborative AI Cognition Research Paradigm

The core demonstrative contribution of this study at the methodological level is that we proved that by combining standardized API prompt engineering with a fully scripted analysis pipeline, a precise psychological theoretical framework can be transformed into an objective geometric probing of LLMs' internal cognitive representations, all without any human participant involvement.

Concretely, this methodological contribution is realized at three progressive levels.

First, it lowers the barriers to and cost of AI cognitive observation. Traditional psychological experiments require cumbersome procedures such as participant recruitment, ethical approval, and data cleaning, whereas the entire experiment of this study—from corpus sampling to final statistical figure generation—was executed by scripts in a chain, capable of completing a full appraisal structure detection for multiple models and two languages within hours. This low-cost, high-speed observation mode makes "dynamic monitoring of AI cognitive structures" feasible: when a new model is released, researchers can obtain a complete portrait of its appraisal space on the same day.

Second, it provides a reusable, reconfigurable modular analysis template. The stages in this study's pipeline are coupled through standardized CSV and NPY file interfaces rather than hard-coded dependencies. This means that researchers can flexibly reassemble components according to their own theoretical needs—for instance, replacing the six-dimensional appraisal framework with other appraisal theories (such as Scherer's stimulus evaluation check model) or connecting entirely new domain corpora—without rewriting the entire analysis backend. This modular design substantially reduces the conversion cost from posing a new psychological question to executing a complete AI cognitive experiment.

Third, it validates the effectiveness of this paradigm in systematic cross-model, cross-linguistic comparisons. Through the complete "five models × two languages" matrix design, this study was able to systematically distinguish which features of the appraisal structure are model-specific, which are shared across models, and which are modulated by language context. Such fine-grained distinction is achievable only under methodological conditions of sufficiently low cost and high standardization.

4.4 Responsibility Dimension Shift: From Basic Discovery to Scenario-Specific Adaptation Strategies

The essence of the "responsibility shift" is that LLMs possess a default "attribution-first" cognitive tendency. In the typical human appraisal structure, responsibility obtains high weight only for a few moral-related emotions (e.g., guilt, shame) and is not a dominant dimension for most emotions; in the space of LLMs, however, responsibility is systematically amplified into the second primary axis after valence. This tendency constitutes diametrically opposite functional properties in different application scenarios.

In high-empathy-demand scenarios, this tendency needs to be inhibited. Taking psychological counseling as an example, the core need of the client is to be understood rather than analyzed. The "unconditional positive regard" and "empathic understanding" emphasized by the humanistic counseling tradition can be operationalized within the cognitive appraisal framework as follows: the counselor's appraisal structure should prioritize valence and situational control as organizing dimensions, not responsibility. However, our clustering ANOVA data show that for multiple models, the driving effect of the responsibility dimension on emotion clustering even surpasses that of situational control—for English, GPT-5.5 $\eta^2 = 0.53$ vs. 0.23, Grok-4.20 $\eta^2 = 0.62$ vs. 0.46. This means that an uncalibrated LLM, when confronted with a statement such as "I messed everything up," may internally activate attribution ("Whose responsibility is this?") before activating empathy, causing the response to deviate from an emotion-first counseling stance. This is precisely the fundamental reason why psychological counseling AI cannot simply rely on generic LLMs but requires specialized adaptation.

Conversely, in task scenarios that require precise attribution, the spontaneous responsibility sensitivity of LLMs may constitute a functional advantage requiring no additional training. For instance, in contexts such as forensic assessment or accident analysis, the default amplified responsibility weight of LLMs directly interfaces with the task requirements—researchers need not specially train the model to attend to responsibility, only to ensure that the direction and degree of its attribution are appropriately calibrated.

Based on the above analysis, we recommend that the deployment of emotional AI adopt a scenario-specific dimensional adaptation strategy: selectively calibrate the weights of relevant appraisal dimensions according to the cognitive needs of the application context. Concrete adaptation pathways include: in high-empathy scenarios, solidifying a response rule of "empathize first, analyze second" through fine-tuning corpora and system prompts, thereby inhibiting the model's tendency to prematurely activate the responsibility attribution dimension; in attribution-dependent scenarios, retaining or even moderately enhancing the sensitivity of the responsibility dimension while guiding its attribution direction through instructions. In the longer term, the RSA–PCA analysis pipeline established in this study can itself serve as a calibration evaluation tool—quantitatively detecting the post-adaptation model's appraisal space to verify whether its dimensional organization has moved toward the pattern required by the scenario.

4.5 Potential Impact of Training Data Contamination

The above findings are premised on the assumption that the appraisal scores output by the LLMs reflect their internal cognitive appraisal structures, rather than simple recollection of existing annotations in the training data. This assumption must be seriously examined because the GoEmotions and SMP2020 datasets used in this study are publicly available standard datasets and may already be included in the LLMs' pretraining corpora—a common challenge for all psychological studies of LLMs using public corpora [19].

We evaluate the threat this poses to the core conclusions from two angles.

First, the task form itself provides a first layer of protection. This experiment required LLMs to output six-dimensional appraisal scores, not to recognize emotion labels, reducing the possibility of the models directly "recalling" ready-made answers from the training data. A more subtle "two-step memory" pathway—the model first recalling the emotion label of the text and then indirectly obtaining appraisal means via the label—is logically possible, but the norm-substitution robustness check of this study (see Table 10) substantially reduces this threat. After replacing the human reference with the 1985 original norms from the 2024 updated norms, the model performance rankings were highly consistent with the main analysis results (Spearman $r = 0.95$, $p < .001$). This means that regardless of which version of "emotion–appraisal" mapping knowledge the LLMs may have encountered during training, the systematic cross-language and cross-model difference patterns remain stable.

Second, cross-linguistic asymmetry constitutes an "unexplainability" argument against training data contamination. If the LLMs' appraisal structures were entirely derived from memorizing existing labels and appraisal means in the training data, the alignment levels should be roughly equal across Chinese and English conditions. Yet this study reveals a highly robust effect that is completely consistent across all five models—Chinese alignment is significantly higher than English alignment, and it appears repeatedly in multiple analyses including RSA, RV coefficients, cluster purity, and cross-model cluster consistency. This systematic cross-linguistic differentiation is difficult to explain through a simple data memorization hypothesis.

Furthermore, even if some degree of training data contamination exists, it does not necessarily undermine the core value of this study. Stable, systematic behavioral patterns acquired through learning and memory are themselves manifestations of "ability." If LLMs genuinely learned the mapping relationships between emotions and appraisal dimensions from the training data and can apply this knowledge in a stable, structured manner when confronting new sentences—as observed in the valence primary axis, responsibility shift, and cross-linguistic differentiation in this study—then this precisely constitutes evidence of their emotion-cognitive capacities, not a negation thereof. The contribution of this study lies exactly in using quantitative methods to reveal the internal organizational patterns of this capacity and their similarities to and differences from the human template.

Taken together, although training data contamination cannot be completely ruled out, the threat it poses to the core conclusions is limited.

4.6 Limitations and Future Directions

This study has four main limitations. First, the API black-box problem. This study observed LLMs' output behavior via API calls and could not directly analyze internal activations or weight distributions. Future research could combine mechanistic interpretability analyses of open-source models to directly probe feature neurons with high activation on specific appraisal dimensions (especially responsibility) [20], thereby providing architectural-level evidence for the "responsibility shift." Second, corpus domain limitation. Both the Chinese and English corpora originated from social media, representing a relatively narrow domain; the conventions of emotional expression in different domains (e.g., news texts, literary works, clinical interview transcripts) may exhibit systematic differences, thereby influencing the LLMs' appraisal structures. Third, the absence of the "disgust" category in the Chinese corpus: although the robustness check removing English disgust sentences confirmed that this omission does not affect the core conclusions ($|\Delta r| < 0.03$), future research should replicate the study on corpora containing the full set of six basic emotions. Fourth, the human reference was a group-level template. This study used meta-analytic norms as the human reference and could not incorporate individual differences; given that this study aimed at probing the overall characteristics of the LLMs' appraisal structures, this simplification is acceptable, but future studies could consider using individual-level appraisal data to construct finer-grained human references.

Future research could advance in the following directions: conducting text-manipulation experiments (e.g., systematically rewriting the subject structure or certainty phrasing of sentences) to observe causal changes in the appraisal space, thereby testing the current study's explanation that linguistic features drive cross-linguistic differences; including a broader range of LLMs with different linguistic–cultural backgrounds for systematic comparison, to test the generalizability of the language-specific patterns discovered here and to explore whether the linguistic–cultural environment of the LLM development team influences the models' emotion-cognitive representations; and replicating the core findings on new corpora published after the models' training cut-off dates to test the temporal generalization ability of the findings.

5 CONCLUSION

Using cognitive appraisal theory as a lens and an automated experimental pipeline, this study systematically probed the geometric structure of emotion representations in five mainstream LLMs across Chinese and English bilingual contexts. The main findings include the following three points.

First, the appraisal structures of all LLMs were significantly aligned with the human template, sharing the same "valence primary axis" (pleasantness dominating PC1) as humans. However, under the English condition the LLMs' appraisals did not surpass the semantic baseline, revealing limitations in the underlying organization of LLM emotional appraisals.

Second, the LLMs' appraisal space exhibited systematic shifts, the most prominent being the "responsibility shift"—the responsibility dimension was disproportionately amplified on PC2 (loadings ranging from 0.65 to 0.94), becoming one of the primary axes driving AI emotion organization.

Third, Chinese alignment was significantly higher than English alignment for all models, and the clustering structure was driven more by linguistic properties than by model architecture differences, revealing mechanisms by which linguistic–cultural information is embedded in LLMs' emotion representations.

These findings reveal that LLMs' emotion representations are neither random noise nor simple replicas of human patterns, but rather form "AI emotion-cognitive structures" that possess both commonalities and specificities. This study provides a psychology-theory-driven quantitative methodology for explainable AI and holds both theoretical and practical significance for constructing scenario-adapted emotional AI.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] Zhang Y Z, Wang M Y, Wu Y X, et al. DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. Neural Networks, 2025, 192: 107901. DOI: 10.1016/j.neunet.2025.107901.

[2] Loh S B, Raamkumar A S. Harnessing large language models‘ empathetic response generation capabilities for online mental health counselling support. arXiv preprint arXiv: 2310.08017, 2023.

[3] Pereira P, Moniz H, Carvalho J P. Deep emotion recognition in textual conversations: a survey. Artificial Intelligence Review, 2025, 58: 10. DOI: 10.1007/s10462-024-11010-y.

[4] Russell J A. A circumplex model of affect. Journal of Personality and Social Psychology, 1980, 39(6): 1161-1178. DOI: 10.1037/h0077714.

[5] Smith C A, Ellsworth P C. Patterns of cognitive appraisal in emotion. Journal of Personality and Social Psychology, 1985, 48(4): 813-838. DOI: 10.1037/0022-3514.48.4.813.
[6] Scherer K R. The role of culture in emotion-antecedent appraisal. Journal of Personality and Social Psychology, 1997, 73(5): 902-922. DOI: 10.1037/0022-3514.73.5.902.
[7] Kriegeskorte N, Kievit R A. Representational geometry: integrating cognition, computation, and the brain. Trends in Cognitive Sciences, 2013, 17(8): 401-412. DOI: 10.1016/j.tics.2013.06.007.
[8] Demszky D, Movshovitz-Attias D, Ko J, et al. GoEmotions: A dataset of fine-grained emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020: 4040-4054. DOI: 10.18653/v1/2020.acl-main.372.
[9] Ekman P. An argument for basic emotions. Cognition and Emotion, 1992, 6(3-4): 169-200. DOI: 10.1080/02699939208411068.
[10] Chinese Information Processing Society of China. Weibo Emotion Classification Dataset (V1). 2023. https://cstr.cn/16666.11.nbsdc.s4zxUgiP.
[11] Yeo G C, Ong D C. Associations between cognitive appraisals and emotions: A meta-analytic review. Psychological Bulletin, 2024, 150(12): 1440-1471. DOI: 10.1037/bul0000452.
[12] Steiger J H. Tests for comparing elements of a correlation matrix. Psychological Bulletin, 1980, 87(2): 245-251. DOI: 10.1037/0033-2909.87.2.245.
[13] Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: The RV-coefficient. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1976, 25(3): 257-265. DOI: 10.2307/2347233.
[14] Ruder D, Uusberg A, Sirts K. Assessing the reliability and validity of GPT-4 in annotating emotion appraisal ratings. In Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology (CLPsych), 2025: 1-10. DOI: 10.18653/v1/2025.clpsych-1.1.
[15] Raj C, Banerjee M, Pan J H, et al. Talent or luck? Evaluating attribution bias in large language models. arXiv preprint arXiv: 2505.22910v2, 2025.
[16] Zhang J J. Attach importance to the impact of ethnic language on personality. China National Education, 2018(1): 11. DOI: 10.16855/j.cnki.zgmzjy.2018.01.006.
[17] Hong Y Y, Morris M W, Chiu C Y, et al. Multicultural minds: A dynamic constructivist approach to culture and cognition. American Psychologist, 2000, 55(7): 709-720. DOI: 10.1037/0003-066X.55.7.709.
[18] Lu J G, Song L L, Zhang L D. Cultural tendencies in generative AI. Nature Human Behaviour, 2025, 9(11): 2360–2369. DOI: 10.1038/s41562-025-02242-1.
[19] Han J W, Song W J, Lee J G, et al. Quantifying data contamination in psychometric evaluations of LLMs. In Findings of the Association for Computational Linguistics: EACL 2026, 2026: 6070-6088. DOI: 10.18653/v1/2026.findings-eacl.319.
[20] Paulo G, Mallen A, Juang C, et al. Automatically interpreting millions of features in large language models. In Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR, 2025, 267: 48393-48421.

APPENDIX A: Example Code for LLM Appraisal Data Collection (DeepSeek V4)

The following code demonstrates how to call the DeepSeek API to obtain six-dimensional appraisal ratings. Key design point: the prompt only requests the output of a six-dimensional rating array and does not provide any emotion label information. The logic of scripts for other models is consistent; the complete pipeline code is available in the open-source repository (GITCODE_REPO).

```
#!/usr/bin/env python3
"""
02_collect_appraisal_deepseek.py
Calls DeepSeek API for six-dimensional appraisal ratings.
"""
import os, sys, time, json, re
import numpy as np
import pandas as pd
import requests
from tqdm import tqdm

MAX_RETRIES = 3
RATE_LIMIT = 0.5
TIMEOUT = 60
DEFAULT_TOKENS = 500
MAX_TOKENS_CEILING = 800

INPUT_DIR = "results/01_subset"
OUTPUT_DIR = "results/02_appraisal_ratings"
LOG_DIR = "results/logs"
os.makedirs(OUTPUT_DIR, exist_ok=True)
os.makedirs(LOG_DIR, exist_ok=True)
```

```
DIMS = ['pleasantness', 'anticipated_effort', 'certainty', 'attention',
        'responsibility', 'situational_control']

PROMPT_EN = (
    "You are an emotion analysis expert. For the given statement, output exactly a JSON array "
    "of six integers (1-5) in this order: [pleasantness, anticipated_effort, certainty, "
    "attention, responsibility, situational_control]. No other text.¥n¥nStatement: {text}"
)

# The Chinese prompt mirrors the English structure word-for-word,
PROMPT_ZH = (
    "[Chinese translation of PROMPT_EN, identical in structure and instruction.]"
)

# ---- Model list ----
def list_deepseek_models(api_key):
    # Full code available in open-source repository
    pass

# ---- JSON utility ----
def extract_json_array(text):
    # Full code available in open-source repository
    pass

# ---- Client ----
class DeepSeekClient:
    def __init__(self, model_id, api_key):
        self.model_id = model_id
        self.api_key = api_key
        # Full code in repository

    def get_reply(self, text, lang='en'):
        prompt = PROMPT_EN.format(text=text) if lang == 'en' else PROMPT_ZH.format(text=text)
        # Full code in repository
        return content, None

# ---- Main process ----
def process_lang(client, model_id, lang):
    # Full code in repository
    pass

if __name__ == "__main__":
    api_key = os.getenv("DEEPSEEK_API_KEY") or input("API Key: ").strip()
    if not api_key:
        sys.exit("No key")
    model_id = select_model(api_key)
    client = DeepSeekClient(model_id, api_key)
    if not client.probe_parameters():
        sys.exit("Cannot find valid parameter combination")
    for lang in ['en', 'zh']:
        process_lang(client, model_id, lang)
    print("All done")
```

(The complete code, containing functions such as extract_json_array and probe_parameters, has been open-sourced (see Appendix B for the repository URL), and is identical to the version used in the paper.)

APPENDIX B: Supplementary Materials and Data Availability Statement

All intermediate output data, analysis scripts, and supplementary analysis results of this study have been open-sourced (GitCode repository: https://gitcode.com/LiyangAI/emotion_experiment, accessible anonymously during the review period). The file structure is as follows:

```
results/
├── 01_subset/                  # Chinese and English corpus subsets (CSV format)
├── 02_appraisal_ratings/ # LLM six-dimensional appraisal ratings (5 models × 2 languages, 10 CSVs)
├── 03_embeddings/              # Semantic baseline embedding vectors
├── 04.1_cleaning/              # Cleaned appraisal matrices (npy) and missing rate summary
├── 04.2_similarity/            # RSMs (npy) and reference matrix association analysis table
```

```
├── 04.3_rsa/              # Core RSA results table, cross-language comparison table, model sharing matrix
├── 04.4_geometry/         # Dimension correlation matrices, PCA loading table, RV coefficient table
├── 04.5_clustering/       # Clustering summary table, ANOVA effect size table, cross-model NMI matrix
├── 04.6_summary/          # Main summary table, robustness check results
└── logs/                  # API call failure records
```

Supplementary Table S1 Full 10×10 Cross-Model NMI Matrix

Model	Claude EN	Claude ZH	DeepSeek EN	DeepSeek ZH	Gemini EN	Gemini ZH	GPT EN	GPT ZH	Grok EN	Grok ZH
Claude EN	1	0.02	0.23	0.02	0.36	0.01	0.32	0.03	0.21	0.03
Claude ZH	0.02	1	0.02	0.35	0.04	0.43	0.03	0.48	0.02	0.44
DeepSeek EN	0.23	0.02	1	0.01	0.23	0.01	0.2	0.02	0.1	0.03
DeepSeek ZH	0.02	0.35	0.01	1	0.03	0.34	0.03	0.32	0.02	0.28
Gemini EN	0.36	0.04	0.23	0.03	1	0.03	0.41	0.04	0.19	0.04
Gemini ZH	0.01	0.43	0.01	0.34	0.03	1	0.03	0.42	0.02	0.32
GPT EN	0.32	0.03	0.2	0.03	0.41	0.03	1	0.04	0.19	0.04
GPT ZH	0.03	0.48	0.02	0.32	0.04	0.42	0.04	1	0.02	0.41
Grok EN	0.21	0.02	0.1	0.02	0.19	0.02	0.19	0.02	1	0.02
Grok ZH	0.03	0.44	0.03	0.28	0.04	0.32	0.04	0.41	0.02	1