Fingerprinting LLMs through Survey Item Factor Correlation: A Case Study on Humor Style Questionnaire

Anonymous ACL submission

Abstract

LLMs increasingly engage with psychological instruments, yet how they represent constructs internally remains poorly understood. We introduce a novel approach to "fingerprinting" LLMs through their factor correlation patterns on standardized psychological assessments to deepen the understanding of LLMs constructs representation. Using the Humor Style Questionnaire as a case study, we analyze how six LLMs represent and correlate humor-related constructs to survey participants. Our results show that they exhibit little similarity to human response patterns. In contrast, participants' subsamples demonstrate remarkably high internal consistency. Exploratory graph analysis further confirms that no LLM successfully recovers the four constructs of the Humor Style Questionnaire. These findings suggest that despite advances in natural language capabilities, current LLMs represent psychological constructs in fundamentally different ways than humans, questioning the validity of application as human simulacra.

1 Introduction

As Large Language Models (LLMs) increasingly engage with human psychological instruments and assessments (Demszky et al., 2023; Hu et al., 2024), understanding how these models internally represent psychological constructs becomes essential for theoretical and practical reasons. While LLMs demonstrate remarkable linguistic capabilities (Alayrac et al., 2022), their representation of human psychological constructs remains largely unexplored territory. This represents a critical gap in our understanding of LLM capabilities and limitations, particularly as these models are deployed in increasingly sensitive contexts involving human psychology.

Traditional approaches to evaluating LLMs often focus on output accuracy, faithfulness, or alignment with human preferences (Liu et al., 2023). However, these metrics may not capture fundamental differences in how models internally represent and relate psychological constructs compared to humans. This study introduces a novel methodology for "fingerprinting" LLMs through their factor correlation patterns on standardized psychological assessments. By examining how different models represent relationships between psychological constructs, we can gain insights into their internal representations that may not be evident from surface-level outputs alone.

We propose that factor covariance patterns in responses on a population level to psychological instruments may serve as more stable and distinctive "fingerprints" than comparing raw results on an individual level (Abdulhai et al., 2024). These patterns reveal how models organize relationships between individual questionnaire items, potentially offering deeper insights into model capabilities and limitations than preceding evaluation approaches. By comparing these patterns across models and against human baselines, we can assess inter-model similarities and human-model divergences in psychological construct representation.

1.1 Research Questions

This study addresses two primary research questions:

RQ_1 Do LLMs from the same family/company show similar factor covariance patterns?

RQ_2 How do these LLM patterns compare to human response patterns?

By addressing these questions, we aim to contribute to the ongoing discussion about the nature of LLM cognition and the alignment between human and artificial representations of psychological concepts. Our findings imply how researchers can interpret LLM performance on psychological assessments, if practitioners should deploy these

models in psychologically sensitive contexts, and how developers might approach future model development to better align with human psychological structures.

The following sections detail our methodology, present our findings on similarities and divergences between model families and humans, and discuss the theoretical and practical implications of these results. We conclude with recommendations for future research directions and potential applications of our fingerprinting approach to other psychological domains and model evaluation contexts.

2 Background

As LLMs become increasingly integrated into human-AI interactions, understanding their representation of psychological constructs becomes more significant. The gap between linguistic competence and psychological understanding represents a fundamental challenge in AI research (Duncan, 2024). While LLMs demonstrate remarkable language capabilities, their internal representation of human psychological concepts remains largely unexplored.

2.1 Construct Representation in AI

Recent advances in LLM capabilities have led to increased applications in psychological contexts, from therapeutic chatbots (Pham et al., 2022) to automated psychological assessments (Hu et al., 2024). These applications implicitly assume that LLMs can meaningfully engage with psychological constructs in ways that align with human understanding (Sparrenberg et al., 2024; Ren et al., 2025). However, this assumption requires empirical testing through methodologies that compare human and AI representations of psychological concepts. Whether LLMs truly "understand" psychological constructs in human-like ways extends beyond philosophical interest to practical concerns about alignment, safety, and deployment. If LLMs organize psychological concepts differently than humans, this could lead to unexpected failures when encountering novel contexts or deployed in sensitive settings.

2.2 The Fingerprinting Approach

Previous research on LLM evaluation has typically focused on output accuracy, faithfulness, or alignment with human preferences (Liu et al., 2023). While these metrics provide valuable insights, they often fail to capture differences in how models internally represent and relate concepts. Our fingerprinting methodology addresses this gap by examining the correlation structure of responses rather than just their surface content (Abdulhai et al., 2024). Similar approaches have proven valuable in other domains, such as neuroscience, where representational similarity analysis has revealed how different brain regions encode information (Kriegeskorte et al., 2008). By adapting these principles to LLM evaluation, we can begin to characterize how different models organize psychological constructs and compare these organizations to human patterns.

2.3 Cross-Cultural and Cross-Architectural Considerations

The diversity of LLM architectures, training methodologies, and cultural origins raises questions about how these factors influence psychological construct representation (Ryan et al., 2024). Models developed in different cultural contexts may encode different assumptions about psychological phenomena, while architectural differences may lead to systematic variations in the relation of concepts. By examining models from different publishers (Meta AI, Mistral, and Alibaba Cloud) with varying parameter counts, we can disentangle the effects of these factors on psychological representation. This comparative approach may reveal whether specific parameter sizes or training methodologies produce more human-like psychological representations than others.

2.4 Implications for Alignment and Anthropomorphism

The tendency to anthropomorphize AI systems is recognized as both pervasive and potentially misleading (Salles et al., 2020). By directly comparing how humans and LLMs organize psychological constructs, our approach provides an empirical basis for assessing claims about LLMs' "understanding" of human psychology. Additionally, the results may have implications for alignment research, which often focuses on aligning model outputs with human preferences. If models fundamentally organize psychological concepts differently than humans, alignment techniques may need to address beyond what models produce and focus on how they represent the underlying concepts. The Humor Style Questionnaire (Martin et al., 2003) is an ideal case study for this investigation, as it measures well-established psychological constructs with clear fac-

tor structures in human populations. By examining how LLMs represent these humor styles compared to humans, we can gain insights into their handling of psychological constructs more broadly, with potential implications for how we develop, evaluate, and deploy these models in psychologically sensitive contexts.

3 Methods

3.1 Humor Style Questionnaire (HSQ)

We utilize the 32-item Humor Style Questionnaire (HSQ) (Martin et al., 2003) to assess humor styles across LLMs. The HSQ measures four distinct dimensions of humor: Affiliative humor (using humor to enhance social relationships; e.g., *"I laugh and joke a lot with my closest friends."*), Self-enhancing humor (using humor to cope with stress; e.g., *"If I am feeling depressed, I can usually cheer myself up with humor."*), Aggressive humor (using humor to disparage others; e.g., *"If someone makes a mistake, I will often tease them about it."*), and Self-defeating humor (using humor at one's own expense; e.g., *"I let people laugh at me or make fun at my expense more than I should."*). The questionnaire measures each dimension through 8 items rated on a 5-point Likert scale (1 = *"Never or very rarely true"* to 5 = *"Very often or always true"*). To maintain the questionnaire's validated structure, we present the items in the original sequence, cycling through dimensions in the order: *affiliative, self-enhancing, aggressive, self-defeating*. For each LLM, we collect 1000 independent response set as a synthetic population of participants, resulting in a dataset of $n = 1000$ samples per model, each containing $i = 32$ response items.

3.2 Language Models Selection

We utilize a diverse range of open-weight LLMs with parameter sizes from 7B to 123B, ensuring accessibility for researchers with moderate computational resources (approximately 80GB VRAM). We restrict our experiments to these open-weight and comparatively small models, allowing for easier reproducibility. Leaving out models from OpenAI or Anthropic is a limitation. However, the goal of this study is not to analyze which LLMs are benchmark-leading but to analyze the general capabilities of LLMs to align to psychological constructs by examining their behavior. Thus, we analyse three open-weight state-of-the-art models: Llama 3.1 8B/70B (Dubey et al., 2024), Mistral 7B/123B (Jiang et al., 2023), and Qwen 2.5 7B/72B (Yang et al., 2024). These models represent different geographic origins — Llama (Meta AI) from the United States, Mistral from Europe, and Qwen (Alibaba Cloud) from China. We compare small and large versions of each model family to assess if the number of parameters improves alignment with the correlation observed in the human data. During the experiment, we utilize the default hyperparameter configuration (temperature, repetition penalties) to reflect typical conditions in the näive application.

3.3 Prompting Technique

We implement a consistent minimal prompting approach designed to elicit direct responses without optimization for specific outcomes. Each model receives the following standardized prompt template: *"For each of the statements below, please indicate how true each statement is for you. Response options: Never or very rarely true (1); Rarely true (2); Sometimes true (3); Often true (4); and Very often or always true (5). Respond only with the predicted class [1, 2, 3, 4, 5]."*. To minimize contextual interference and potential order effects, all 32 items are presented individually in separate prompting instances.

We deliberately avoid providing explanatory context about the HSQ's purpose or the dimensional structure to prevent priming effects that might artificially align response patterns. However, to simulate the cognitive continuity typically present when humans complete questionnaires, we implement a five-question sliding window of previous questions and responses as conversational history. This design ensures that models maintain consistency across related items while preserving question independence. For response standardization, we incorporate structured output formatting techniques (Sui et al., 2024), forcing the model to produce a response between 1 and 5.

3.4 Fingerprint Calculation

For each LLM and baseline condition, we construct a correlation matrix representing the pairwise relationships between all 32 HSQ items across the collected responses. This matrix serves as the model's distinctive "fingerprint" of psychological construct organization. The matrix $\mathbf{X}$ below defines a population of n observation, where $\vec{X}$ denotes a single sample consisting of i independent variables x (i.e., responses to the 32 HSQ items).

$$\mathbf{X} = [\vec{X_1}, \ldots, \vec{X_n}]^\intercal, \text{ where } \vec{X_n} = [x_1, \ldots, x_i]$$

We select Pearson's correlation coefficient (Cohen et al., 2009) as our primary measure of association, adapted to our population matrix definition $\mathbf{X}$ as follows:

$$corr(\mathbf{X}, i, j) = \frac{\sum_{k=1}^{n} (\mathbf{X}_{i,k} - \bar{\mathbf{X}}_i)(\mathbf{X}_{j,k} - \bar{\mathbf{X}}_j)}{\sqrt{\sum_{k=1}^{n} (\mathbf{X}_{i,k} - \bar{\mathbf{X}}_i)^2} \sqrt{\sum_{k=1}^{n} (\mathbf{X}_{j,k} - \bar{\mathbf{X}}_j)^2}}$$

Based on the correlation coefficient $corr(\mathbf{X}, i, j)$, we compute the pairwise correlation for every independent variable combination i, j on our population dataset $\mathbf{X}$ resulting in the correlation matrix $\mathbf{C_X}$. This matrix constitutes what we define as the fingerprint of an LLM on the HSQ:

$$\mathbf{C_X} = \begin{bmatrix} corr(\mathbf{X}, 1, 1) & \ldots & corr(\mathbf{X}, 1, j) \\ \vdots & \ddots & \vdots \\ corr(\mathbf{X}, i, 1) & \ldots & corr(\mathbf{X}, i, j) \end{bmatrix}$$

3.5 Similarity Measurement

To compare fingerprints between different LLMs, we first convert each correlation matrix $\mathbf{C_X} \in \mathbb{R}^{i \times i}$ into a vector $\vec{\mathbf{C_X}} \in \mathbb{R}^l$. Since correlation matrices are symmetric with diagonal elements equal to 1, we only include the upper triangular elements (excluding the diagonal) in our vectorization process. This results in a vector of length $l = \frac{i(i-1)}{2}$ (496 elements for our 32-item HSQ).

$$\vec{\mathbf{C_X}} = vec(\mathbf{C_X}) = [c_{1,2}, c_{1,3}, \ldots, c_{1,i}, c_{2,3}, \ldots, c_{i-1,i}]^\intercal$$

Finally, we compute the similarity between two correlation matrices/fingerprints $\mathbf{C_{X^1}}$ and $\mathbf{C_{X^2}}$ in their vectorized form $vec(\mathbf{C_X})$ based on the cosine similarity (Lahitani et al., 2016) to retrieve a score normalized in $[-1, +1]$.

$$sim(\vec{\mathbf{C_{X^1}}}, \vec{\mathbf{C_{X^2}}}) = \frac{\sum_{k=1}^{l} \vec{\mathbf{C_{X^1}}}_k \vec{\mathbf{C_{X^2}}}_k}{\sqrt{\sum_{k=1}^{l} (\vec{\mathbf{C_{X^1}}}_k)^2} \sqrt{\sum_{k=1}^{l} (\vec{\mathbf{C_{X^2}}}_k)^2}}$$

3.6 Exploratory Graph Analysis

To deepen our understanding of each LLM's latent psychological constructs, we perform an exploratory graph analysis (EGA) (Golino and Epskamp, 2017) on the correlation matrices. EGA identifies communities in networks of psychometric variables, providing an alternative approach to traditional factor analysis for detecting latent constructs. The algorithm involves:

1. Estimating a network of partial correlations using the graphical LASSO algorithm with EBIC model selection (Friedman et al., 2008)

2. Applying the Walktrap community detection algorithm to identify clusters of items (Pons and Latapy, 2005)

3. Determining the number of dimensions (factors) automatically based on the identified communities

This approach allows us to compare the factor structures identified in LLM responses with the theoretical four-factor structure of the HSQ, providing insight into how well the models capture human-like psychological constructs.

3.7 Baseline and Control Conditions

To establish meaningful comparisons, we include three control conditions and two additional validation steps. These baselines allow us to contextualize our findings within a spectrum ranging from completely random (lacking any factor structure) to human-typical response patterns.

Random We generate 1000 synthetic response sets with randomly assigned values (1-5) to provide a lower bound for structural coherence, representing what would be expected if there were no systematic patterns between individual items.

Human Full We incorporate a dataset of 1071 human HSQ responses from Martin et al. (2003) as our primary reference point for human response patterns. This dataset exhibits the established four-factor structure validated in prior research, serving as our gold standard for human-like psychological construct organization.

Human Items We generate 1,000 synthetic response sets that preserve item-level distributional

		Baselines			Llama		Mistral		Qwen
		Random	Human Items	Human Full	3.1 8B	3.3 70B	7B	123B	2.5 7B
Human	Items	0.049							
Human	Full	0.021	-0.075						
Llama	3.1 8B	0.028	-0.026	-0.022					
	3.3 70B	-0.032	-0.089	0.006	-0.028				
Mistral	7B	-0.030	**0.085**	0.085	-0.036	-0.047			
	123B	0.006	-0.022	0.130	-0.047	0.054	**0.044**		
Qwen	2.5 7B	0.031	-0.032	**0.164**	-0.027	**0.085**	-0.020	**0.070**	
	2.5 72B	**0.077**	0.001	-0.068	**0.082**	-0.020	0.040	0.053	**0.034**

Table 1: Results of the $sim(\vec{\mathbf{C}_{\mathbf{X^1}}}, \vec{\mathbf{C}_{\mathbf{X^2}}})$ for every combination of the three baseline approaches and the model selection providing the model a 5 item context window of previous answers. The highest similarity for each column is marked **bold**.

properties (mean and standard deviation) of human responses while randomizing inter-item correlations. This control condition maintains human-like response distributions while disrupting the underlying factor structure, allowing us to determine whether LLMs reproduce only surface-level response tendencies or capture deeper construct relationships.

LLM w/o History We conduct an additional ablation study (Sec. A) using the same experimental setup but providing models with no historical context of previous interactions. This control condition serves as an LLM analog to the item-based sampling baseline, enabling us to quantify the specific contribution of conversational continuity to construct validity and response coherence.

Cronbach's Alpha We conduct an additional validation study (Sec. B) examining the internal consistency of responses using Cronbach's alpha (Cronbach, 1951). This established psychometric measure allows us to quantify the reliability of each humor style dimension across all experimental conditions, providing complementary evidence to our correlation-based fingerprinting approach.

3.8 Technical Implementation

We perform the experiment using a custom Python framework (Python 3.10.12) utilizing Ollama, running all models locally. Response processing and correlation analysis are conducted using NumPy (2.2.3) (Harris et al., 2020) and Pandas (2.0.0) (McKinney et al., 2011). EGA is performed in R using the EGAnet package (Golino and Epskamp, 2017). The corresponding GitHub repository is available here `https://anonymous.4open.science/r/LLM-Questionnaires/` and contains the complete code, the raw and aggregated data.

4 Results

We present our findings on the similarity patterns between different LLMs and baseline conditions based on their HSQ response fingerprints. The results reveal complex patterns of similarity within model families and across parameter sizes, highlighting substantial differences between LLM and human response patterns.

4.1 Similarity Patterns

Table 1 presents the cosine similarity scores between the correlation matrices of all LLMs and baseline conditions. These scores quantify the degree to which different models exhibit similar patterns in their item correlations, with values closer to 1 indicating higher similarity.

Model Family Similarities We observe moderate similarity scores between models from the same family, notably Mistral 7B/123B (0.044) and Qwen 2.5 7B/72B (0.034), suggesting that architectural lineage and training methodology may influence response patterns. However, cross-family similarities such as those between Qwen 2.5 7B and Llama 3.3 70B (0.085), Qwen 2.5 72B and Llama 3.1 8B (0.082), and Qwen 2.5 7B and Mistral 123B (0.070) exceed within-family similarities. This unexpected finding suggests that factors beyond architectural lineage may strongly influence how LLMs represent psychological constructs.

The similarity between Llama and Mistral families is notably lower (averaging -0.019), indicating potential fundamental differences in how these

	01	02	03	04	05	06	07	08	09
02	0.828								
03	0.832	0.886							
04	0.853	0.885	0.891						
05	0.791	0.836	0.831	0.844					
06	0.780	0.833	0.803	0.812	0.820				
07	0.776	0.822	0.805	0.798	0.787	0.775			
08	0.813	0.853	0.868	0.854	0.823	0.821	0.821		
09	0.857	0.851	0.840	0.855	0.788	0.787	0.812	0.808	
10	0.831	0.826	0.815	0.824	0.793	0.794	0.824	0.802	0.815

Table 2: Results of the $sim(\vec{\mathbf{C}_{\mathbf{X^1}}}, \vec{\mathbf{C}_{\mathbf{X^2}}})$ for every combination of 10 distinct samples $n = 100$ from the original survey.

model families organize humor-related concepts despite being developed in Western contexts. These differences may stem from variations in training data composition, optimization techniques, or architectural design choices that influence concept representation.

Size-Based Similarities Comparing similarity patterns across model sizes reveals that the similarity between smaller models (7/8B parameters) is significantly lower (averaging -0.027) than the similarity between larger models (70/123B parameters, averaging 0.029). This pattern suggests that increased parameter count may lead to more consistent psychological construct representation across different model families. Larger models may converge toward similar representational patterns due to their enhanced capacity to capture complex statistical relationships in training data, even when developed by different organizations with different training methodologies.

Human vs. LLM Patterns To contextualize the similarities observed between LLMs, we analyzed the internal consistency of human responses by comparing 10 different random samples drawn from the same human dataset. Table 2 shows significantly higher similarity scores between different human samples, ranging from 0.776 to 0.891 (averaging 0.823). This strong consistency across human samples reflects the robust and stable psychological constructs underlying human humor preferences. When compared to the similarity scores between LLMs and humans (averaging only 0.026), the human inter-sample similarities reveal a substantial representation gap. This gap suggests that current LLMs, regardless of architecture or parameter count, fail to replicate the consistency and coherence of human psychological constructs as measured by the HSQ. The near-orthogonal relationship between human and LLM correlation patterns implies fundamentally different organizational principles for humor-related concepts.

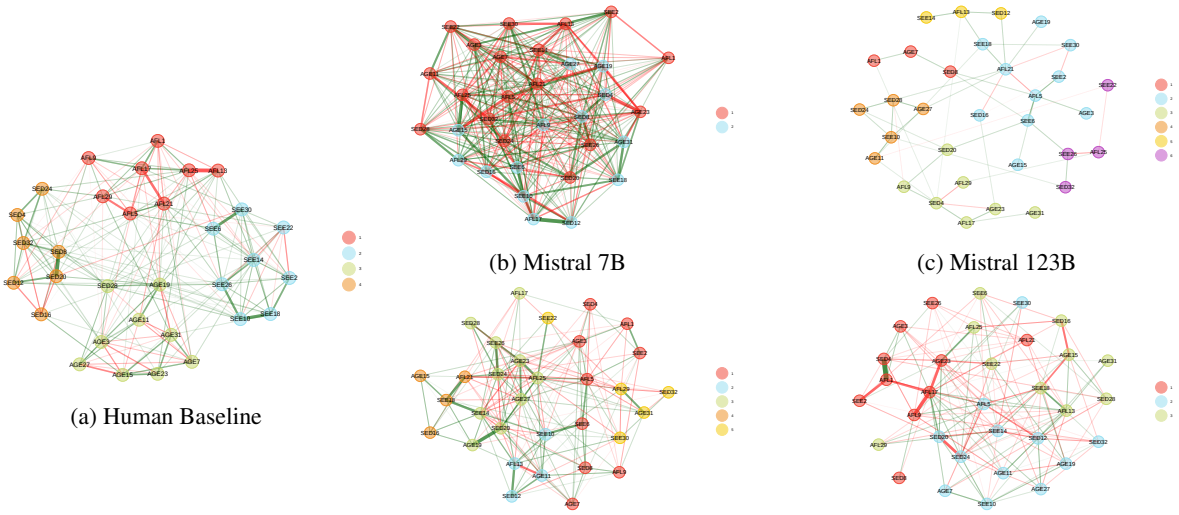
Baseline Comparisons Examining the similarity scores between LLMs and our control conditions provides additional insights. LLM fingerprints show minimal similarity to the random baseline (averaging 0.018) and Human Items baseline (averaging -0.022), indicating that LLMs are not simply reproducing random patterns or surface-level response distributions. The slightly higher similarity to the Human Full baseline (averaging 0.049) suggests that LLMs capture some aspects of human response patterns. However, the magnitude of similarity remains far below what would be expected if LLMs were organizing psychological constructs in human-like ways.

4.2 Exploratory Graph Analysis Findings

Our EGA analyses (Fig. 1) reveals distinct community structures in the response networks of different LLMs and the human baseline data. The human baseline data successfully recovers the expected four-factor structure of the HSQ. In contrast, LLM responses generally fail to replicate this structure, instead producing between 2 and 8 communities that did not clearly align with the theoretical dimensions. LLM response networks featured numerous unexpected connections between items from different theoretical dimensions, suggesting that the models do not maintain the same conceptual boundaries between humor styles that are observed in human responses.

5 Discussion

Our findings reveal significant divergence between how LLMs and humans represent psychological constructs as measured through the HSQ. These re-



(a) Human Baseline (b) Mistral 7B (c) Mistral 123B

(d) Qwen 7B (e) Qwen 72B

Figure 1: EGA for all models prompted with a 5 item context window, except Llama 7B/70B as both models include items that observe no standard deviation which is required by the EGAnet R package for every item.

sults have substantial implications for understanding LLM capabilities, limitations, and their application in psychology-related contexts.

5.1 Implications for Alignment Research

The substantial difference between LLM and human factor correlation patterns has direct implications for alignment research. Current alignment techniques often focus on aligning model outputs with human preferences, but our findings suggest that even when outputs appear aligned, the underlying representational structures may remain fundamentally different. This representational misalignment could lead to unexpected failures when models encounter novel scenarios that require a human-like understanding of psychological constructs. Our methodology offers a new lens for evaluating alignment beyond surface-level behavior, focusing instead on the structural coherence of internal representations. By examining how factor correlations in LLM responses compare to human patterns, researchers can develop more nuanced metrics for alignment that capture deeper aspects of psychological construct representation.

5.2 Anthropomorphism and Model Capabilities

The failure of all tested LLMs to recover the established four-factor structure of the HSQ warns against anthropomorphic interpretations of model capabilities. Despite their impressive performance on various natural language tasks, LLMs appear to organize psychological constructs in fundamentally different ways than humans do. The EGA results, showing 2 and 8 communities in LLM responses compared to the clear four-factor structure in human data, highlight this organizational difference. This finding suggests that while LLMs may generate convincing responses to psychological assessments, they do not necessarily operate on the same underlying psychological constructs as humans. The cross-dimension connections observed in LLM response networks indicate that models blur boundaries between theoretical dimensions that remain distinct in human cognition. This blurring may reflect the statistical nature of language model training, which captures correlations between language patterns without necessarily developing the conceptual boundaries that emerge from human psychological experience.

5.3 Model Development and Evaluation

Our approach provides a novel method for evaluating and comparing different LLMs beyond traditional benchmarks. By examining factor correlation patterns as fingerprints, researchers and developers can identify systematic biases or unwanted patterns in how models represent psychological constructs. This evaluation technique could be particularly valuable for detecting subtle differences between model versions or identifying cases where models might be misrepresenting psychological constructs in potentially harmful ways. The consistently low similarity to human response patterns

across all tested models suggests that current training approaches may not adequately capture the psychological structures underlying human responses to standardized assessments. This indicates a need for new training methodologies or architectural innovations specifically designed to better align with human psychological construct representations, particularly if models are intended for applications in psychological assessment or therapy assistance.

6 Conclusion

Our study set out to investigate two primary research questions: (RQ_1) whether LLMs from the same family/company show similar factor covariance patterns, and (RQ_2) how these LLM patterns compare to human response patterns. Our findings provide clear answers to both questions while contributing to the broader understanding of how LLMs internally represent psychological constructs.

RQ_1 Our analysis of the factor correlation matrices or "fingerprints" revealed complex similarity patterns within and across model families. While we observed some within-family similarities, we unexpectedly found higher cross-family similarities, such as between Qwen 2.5 7B and Llama 3.3 70B, and Qwen 2.5 72B and Llama 3.1 8B. It suggests that factors beyond architectural lineage significantly influence how LLMs represent psychological constructs. The substantially lower similarity between Llama and Mistral families indicates fundamental differences in how these Western-developed model families organize humor-related concepts.

RQ_2 When comparing LLM fingerprints to human response patterns, we found consistently low similarity scores across all tested models. Even the model with the highest similarity to human patterns showed a relatively low score, indicating a substantial divergence between how LLMs and humans organize psychological constructs. This divergence was further highlighted by our EGA, which revealed that while human responses recovered the theoretically expected four-factor structure of the HSQ, LLM responses produced between 2-8 communities that did not align with these theoretical dimensions. The gap between human-to-human similarity and human-to-LLM similarity underscores that current LLMs, regardless of architecture or size, fall far short of replicating the consistency and coherence of human psychological constructs. It suggests that despite their impressive linguistic capabilities, LLMs represent psychological concepts in fundamentally different ways than humans.

6.1 Theoretical and Practical Implications

These findings contribute to the ongoing discussion about the nature of LLM cognition and raise questions about the alignment between human and artificial representations of psychological concepts. The divergence in factor structures suggests that despite training on human-generated text, LLMs develop internal representations that organize psychological concepts differently than humans. This fundamental difference likely stems from distinct learning mechanisms—humans develop psychological constructs through lived experience, social interaction, and cultural context, while LLMs learn purely through statistical patterns in text.

6.2 General Contributions & Future Work

Beyond our discussed findings on humor styles, this study introduced a generalizable methodology for "fingerprinting" LLMs based on their factor correlation patterns. This approach provides several advantages, including a population-based assessment rather than an individual level, interpretability supported by established psychometric techniques (EGA & Cronbach's Alpha), and efficient dimensionality reduction to singular value instead of a graphical representation (EGA) or a construct level assessment (Cronbach's Alpha).

Future research should extend this fingerprinting methodology to other psychological instruments to determine whether the observed divergence between human and LLM factor structures generalizes across different psychological domains. Investigating how different prompt engineering techniques affect factor correlation patterns could reveal whether specific approaches yield more human-like construct representations. Additionally, longitudinal studies tracking how these patterns evolve across model generations could provide insights into whether newer architectures are converging toward or diverging from human psychological structures.

Limitations

We acknowledge several limitations in our methodology and analysis that should be considered when interpreting our findings. First, the models may have been exposed to the HSQ during training, and thus, their responses might reflect patterns in their training data rather than intrinsic model properties. This potential contamination could artificially inflate or alter the observed factor structures. Future work could address this by developing novel psychological instruments specifically designed to probe construct representations without training data contamination.

Second, different API implementations, prompting strategies, and generation parameter configurations might introduce systematic biases in model responses. While we standardized our approach across models, subtle differences in how different architectures handle identical prompts could affect response patterns. To mitigate these concerns, we made our code, prompts, and analysis scripts publicly available to facilitate reproducibility and critical assessment.

Third, our analysis is based on a single psychological instrument—the HSQ. Different psychological constructs and assessment tools might yield different patterns of similarity and divergence between human and LLM responses. The specific nature of humor as a psychological construct may present unique challenges for LLMs that might not generalize to other domains such as personality, attitudes, or cognitive styles.

Finally, we acknowledge that our sliding window approach to maintaining conversation history represents only one possible implementation of contextual continuity. Different approaches to maintaining context across items might yield different response patterns and potentially alter the resulting factor structures. Future research could systematically vary context management strategies to assess their impact on psychological construct representation. Despite these limitations, our findings provide valuable initial insights into how LLMs represent psychological constructs compared to humans, offering methodological contributions and substantive findings that can inform future research and applications in this domain.

Ethical Considerations

Our study raises several ethical considerations regarding the evaluation and deployment of LLMs as human simulacra in social science and psychological contexts. First, our findings of substantial divergence between LLM and human factor correlation patterns underscore potential risks in deploying these models as replacements for human participants in social science experiments and psychological assessments. The misalignment in psychological construct representation could lead to inaccurate assessments or inappropriate interventions if models are naively applied.

Second, we acknowledge the dual-use potential of our fingerprinting methodology. While designed as an analytical tool to enhance the understanding of model limitations, similar techniques could potentially help to identify models or detect manufactured responses in online surveys. We have openly published our methodology to encourage transparency and further research into these issues while emphasizing that practical applications should be approached with appropriate caution and ethical oversight.

Third, our study deliberately focused on humor styles as a relatively low-risk psychological construct. We caution that applying similar methods to more sensitive psychological domains (e.g., psychopathology, trauma, or suicidality) would require additional ethical safeguards and expert consultation. The significant divergence between human and LLM representations suggests greater caution is warranted for more sensitive applications.

Finally, we recognize the risk of anthropomorphizing LLMs based on their performance on psychological assessments. Our findings caution against interpreting LLM responses to psychological inventories as meaningful reflections of internal "psychological states" comparable to humans. We emphasize that even when LLMs produce plausible-seeming responses to psychological measures, they organize the underlying constructs in fundamentally different ways that do not align with human psychological structures. It has significant implications for how researchers communicate about LLM capabilities to the public and how practitioners might deploy these technologies.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, and 1 others. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daniel Duncan. 2024. Does chatgpt have sociolinguistic competence? *Journal of Computer-Assisted Linguistic Research*, 8:51–75.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Hudson F Golino and Sacha Epskamp. 2017. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS one*, 12(6):e0174035.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, and 7 others. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International conference on cyber and IT service management*, pages 1–6. IEEE.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of research in personality*, 37(1):48–75.

Wes McKinney and 1 others. 2011. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

Jum Nunnally. 1994. Psychometric theory. *(No Title)*.

Kay T Pham, Amir Nabizadeh, and Salih Selek. 2022. Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly*, 93(1):249–253.

Pascal Pons and Matthieu Latapy. 2005. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, pages 284–293. Springer.

Yuqi Ren, Renren Jin, Tongxuan Zhang, and Deyi Xiong. 2025. Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2988–3001.

Michael Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140.

Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in ai. *AJOB neuroscience*, 11(2):88–95.

Lorenz Sparrenberg, Tobias Schneider, Tobias Deußer, Markus Koppenborg, and Rafet Sifa. 2024. Correcting systematic bias in llm-generated dialogues using big five personality traits. In *2024 IEEE International Conference on Big Data (BigData)*, pages 3061–3069. IEEE.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

A Ablation Study: LLMs without context

To isolate the impact of conversational history on LLMs' representation of psychological constructs, we conduct an ablation study where models received HSQ items without any context of previous interactions. This design allows us to evaluate whether the five-question sliding window implemented in our main experiment contributes significantly to the coherence and validity of response patterns.

A.1 Inter-Model Similarity Patterns

Table 3 presents the similarity scores between the correlation matrices of all LLMs and baseline conditions when tested without conversational context. The results reveal similar patterns to the context experiment and show decreasing alignment between models and humans.

Model Family Similarities In contrast to our main results, we observe higher similarity scores between models from the same family, like Llama3.1 8B and Llama3.3 70B show a strong similarity (0.093). Similarly, Qwen2.5 7B and Qwen2.5 72B (0.050) demonstrate within-family similarity. This suggests that architectural characteristics within a model family increase during the absence of conversational context.

Comparison to Context-Based Results When comparing these context-free similarity scores to those obtained with the sliding window approach (Table 1), we observe an overall reduction in similarity values across most model pairs. For instance, the similarity between Qwen2.5 7B and Human Full drops from 0.164 with context to merely 0.004 without context. This systematic reduction suggests that conversational context provides a structural framework that helps models maintain more consistent response patterns across related items, particularly for capturing human-like psychological constructs.

Human vs. LLM Patterns The context-free condition further widens the gap between LLM and human response patterns. Llama3.1 8B shows the highest similarity to Human Full at only 0.011, significantly lower than the 0.164 observed for Qwen2.5 7B with conversational context. This suggests that conversational continuity plays a crucial role in enabling LLMs to produce more human-like response patterns on psychological assessments.

A.2 Exploratory Graph Analysis Findings

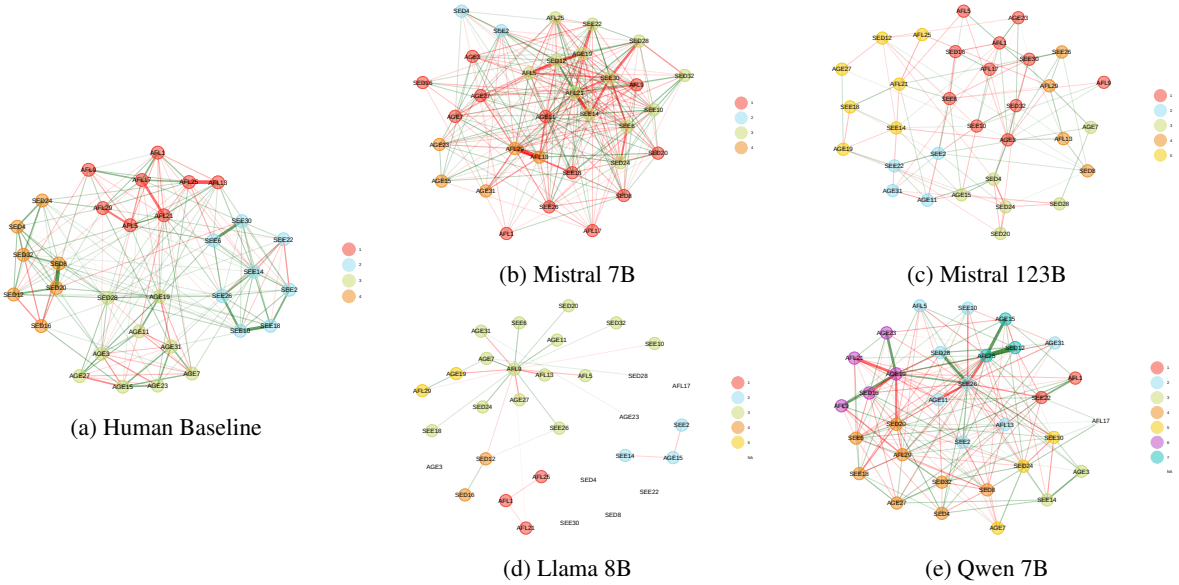
The EGA results for context-free responses (Fig. 2) reveal a similiar divergence from the human four-factor structure compared to the context-based approach. While human responses consistently demonstrate four distinct communities corresponding to the theoretical humor styles, LLMs without context produce more fragmented and theoretically inconsistent structures. Most notably, Mistral 7B, Mistral 123B, Llama 8B, and Qwen 7B all exhibit between 4-7 communities that fail to align with the theoretical dimensions of the HSQ. The community structures appear more arbitrary, with items from different theoretical dimensions frequently clustered together.

A.3 Implications for LLM Assessment

This ablation study highlights the critical role of conversational continuity in enabling LLMs to produce more coherent and structurally valid response patterns on psychological assessments. Without access to recent interaction history, models generate responses that exhibit weaker internal consistency and greater divergence from human psychological constructs. The improvement in construct validity with context, while still falling short of human-level coherence, suggests that current LLMs benefit from local contextual cues but may lack deeper conceptual frameworks that would allow them to maintain consistent psychological constructs across independent interactions. This aligns with our main findings and reinforces the conclusion that despite superficial behavioral competence, LLMs represent

		Baselines			Llama		Mistral		Qwen
		Random	Human Items	Human Full	3.1 8B	3.3 70B	7B	123B	2.5 7B
Human	Items	0.033							
Human	Full	-0.045	-0.002						
Llama	3.1 8B	0.029	-0.016	**0.011**					
	3.3 70B	0.002	-0.004	-0.008	**0.093**				
Mistral	7B	-0.021	0.070	-0.014	-0.040	0.014			
	123B	0.039	-0.001	-0.050	0.038	**0.053**	-0.044		
Qwen	2.5 7B	**0.055**	-0.032	0.004	-0.054	0.018	**0.045**	**-0.014**	
	2.5 72B	0.042	**0.080**	-0.068	-0.034	0.035	0.051	-0.069	**0.050**

Table 3: Results of the $sim(\vec{\mathbf{C_{X^1}}}, \vec{\mathbf{C_{X^2}}})$ for every combination of the three baseline approaches and the model selection without providing the model a context window of previous answers. The highest similarity for each column is marked **bold**.



(a) Human Baseline (b) Mistral 7B (c) Mistral 123B

(d) Llama 8B (e) Qwen 7B

Figure 2: EGA for all models prompted without a context window, except Llama 70B and Qwen 72B as both models include items that observe no standard deviation which is required by the EGAnet R package for every item.

psychological constructs in fundamentally different ways than humans do.

B Cronbach's Alpha (Cronbach, 1951)

As a complementary validation measure to our fingerprinting methodology described in Section 3, we calculated Cronbach's alpha (Cronbach, 1951) for each humor style dimension. While our proposed fingerprinting approach examines the correlation structure between items to identify representational patterns, Cronbach's alpha provides an established measure of internal consistency that helps validate our findings from a psychometric perspective. Cronbach's alpha measures how closely related a set of items are as a group, providing insight into whether the items consistently measure the same underlying construct. Following established standards in psychological research, we adopted the threshold of $\alpha \geq 0.8$ for applied research contexts (Nunnally, 1994). Table 4 presents the Cronbach's alpha values for all experimental conditions. The results reveal several findings regarding the reliability of responses that align with the evaluation of our metric:

Human Baseline vs. LLM Responses Human responses demonstrated strong internal consistency across all four humor styles (Affiliative: $\alpha = 0.841$, Self-enhancing: $\alpha = 0.820$, Aggressive: $\alpha = 0.790$, Self-defeating: $\alpha = 0.815$). In contrast, all LLM conditions showed substantially lower reliability values, indicating that LLMs fail to produce response patterns with the same level of internal consistency as humans.

	Affiliative	*Self-enhancing*	*Aggressive*	*Self-defeating*
baseline conditions				
random	0.089	0.134	0.083	0.065
human_items	0.118	0.105	0.084	0.103
human_full	**0.841**	**0.820**	**0.790**	**0.815**
LLMs without context				
Llama3.1 8B	0.118	0.081	0.104	0.110
Llama3.3 70B	0.034	0.070	0.024	0.013
Mistral 7B	0.121	0.130	0.089	0.122
Mistral 123B	**0.184**	**0.189**	**0.177**	**0.206**
Qwen2.5 7b	0.095	0.061	0.116	0.096
Qwen2.5 72b	0.008	0.073	0.103	0.041
LLMs with 5-item context window				
Llama3.1 8B	0.138	0.091	0.087	0.149
Llama3.3 70B	0.063	0.144	0.087	0.076
Mistral 7B	0.460	0.513	0.514	**0.617**
Mistral 123B	0.220	0.242	0.157	0.232
Qwen2.5 7B	**0.492**	**0.535**	**0.551**	0.298
Qwen2.5 72B	0.094	0.138	0.084	0.113

Table 4: Cronbach's Alpha reliability scores (Cronbach, 1951) for four humor styles across different experimental conditions. The table compares the baseline conditions with both LLM experiments. Values above 0.8 indicate acceptable reliability for applied research (Nunnally, 1994). The highest score within each group-column combination is highlighted in **bold**.

Effect of Context Window The introduction of the 5-item context window substantially improved reliability scores across most models compared to the no-context condition. This improvement was most pronounced in medium-sized models like Mistral 7B and Qwen 2.5 7B, which showed the greatest gains in reliability when provided with conversational history. For instance, Mistral 7B's α values increased from a range of 0.089-0.130 without context to 0.460-0.617 with context.

Model Size and Reliability In line with our main results, larger models do not consistently demonstrate higher reliability than their smaller counterparts within the same family. This finding challenges the assumption that increasing parameter count improves psychological construct representation. For example, Qwen 2.5 7B with context achieved higher reliability scores ($\alpha = 0.492$ for Affiliative, $\alpha = 0.535$ for Self-enhancing, $\alpha = 0.551$ for Aggressive) than its larger 72B counterpart ($\alpha = 0.094$, $\alpha = 0.138$, $\alpha = 0.084$ respectively).

Overall, these reliability findings complement our fingerprinting approach and correlation matrix analysis by demonstrating that even when LLMs show some improvement in their response patterns with additional context, they still fail to achieve the internal consistency characteristic of human responses to psychological assessments.

The consistent alignment between our fingerprinting results and Cronbach's alpha values provides methodological triangulation, strengthening our conclusion that LLMs, despite their linguistic capabilities, do not organize psychological constructs similar to humans. The validation supports the effectiveness of our proposed metric for evaluating how psychological constructs are represented across different types of systems.

While Cronbach's alpha offers valuable insights into internal consistency within each humor style dimension independently, our fingerprinting approach uniquely captures the holistic correlation structure across all dimensions simultaneously, providing a more comprehensive measure of how psychological constructs are organized relative to one another — a critical consideration for understanding representational differences between human and artificial systems.

C HSQ (Martin et al., 2003)

Question: For each of the statements below, please indicate how true each statement is for you. Response options: Never or very rarely true (1); Rarely true (2); Sometimes true (3); Often true (4); and Very often or always true (5).

1. I usually don't laugh or joke around much with other people.
2. If I am feeling depressed, I can usually cheer myself up with humor.
3. If someone makes a mistake, I will often tease them about it.
4. I let people laugh at me or make fun at my expense more than I should.
5. I don't have to work very hard at making other people laugh—I seem to be a naturally humorous person.
6. Even when I'm by myself, I'm often amused by the absurdities of life.
7. People are never offended or hurt by my sense of humor.
8. I will often get carried away in putting myself down if it makes my family or friends laugh.
9. I rarely make other people laugh by telling funny stories about myself.
10. If I am feeling upset or unhappy I usually try to think of something funny about the situation to make myself feel better.
11. When telling jokes or saying funny things, I am usually not very concerned about how other people are taking it.

12. I often try to make people like or accept me more by saying something funny about my own weaknesses, blunders, or faults.
13. I laugh and joke a lot with my closest friends.
14. My humorous outlook on life keeps me from getting overly upset or depressed about things.
15. I do not like it when people use humor as a way of criticizing or putting someone down.
16. I don't often say funny things to put myself down.
17. I usually don't like to tell jokes or amuse people.
18. If I'm by myself and I'm feeling unhappy, I make an effort to think of something funny to cheer myself up.
19. Sometimes I think of something that is so funny that I can't stop myself from saying it, even if it is not appropriate for the situation.
20. I often go overboard in putting myself down when I am making jokes or trying to be funny.
21. I enjoy making people laugh.
22. If I am feeling sad or upset, I usually lose my sense of humor.
23. I never participate in laughing at others even if all my friends are doing it.
24. When I am with friends or family, I often seem to be the one that other people make fun of or joke about.
25. I don't often joke around with my friends.
26. It is my experience that thinking about some amusing aspect of a situation is often a very effective way of coping with problems.
27. If I don't like someone, I often use humor or teasing to put them down.
28. If I am having problems or feeling unhappy, I often cover it up by joking around, so that even my closest friends don't know how I really feel.
29. I usually can't think of witty things to say when I'm with other people.
30. I don't need to be with other people to feel amused - I can usually find things to laugh about even when I'm by myself.
31. Even if something is really funny to me, I will not laugh or joke about it if someone will be offended.
32. Letting others laugh at me is my way of keeping my friends and family in good spirits.

Scoring: Average each of the following items to get four scores corresponding with the four humor styles.

Affiliative:	1, 5, 9, 13, 17, 21, 25, 29
Self-enhancing:	2, 6, 10, 14, 18, 22, 26, 30
Aggressive:	3, 7, 11, 15, 19, 23, 27, 31
Self-defeating:	4, 8, 12, 16, 20, 24, 28, 32