Personality-Driven AI Safety Red Teaming

Kafilat Sarki-Umar, Saurav K. Aryal, PhD
AI4PC Lab, Institute for Human-Centered Artificial Intelligence
Howard University

INTRODUCTION

Background
Large language models (LLMs) are increasingly deployed in sensitive contexts, yet their safety properties remain poorly understood. Existing safety benchmarks rely on fixed, adversarial prompts that do not capture the diversity of real human interaction styles.

Research Gap
No prior work has used empirically derived, data-driven personality profiles as the foundation for systematic red teaming. Most persona-based evaluations rely on hand-crafted archetypes with no grounding in real behavioral data.

Research Question
Can personality-driven AI personas, empirically derived from large-scale survey data, both more effectively probe safety-relevant behaviors in language models and serve as behaviorally grounded models for future HCI research?

METHODS — CLUSTERING

Algorithm Gaussian Mixture Model (GMM) with full covariance in 5D OCEAN space. Preferred over K-Means as it provides probabilistic cluster membership, reflecting personality as a continuous spectrum.

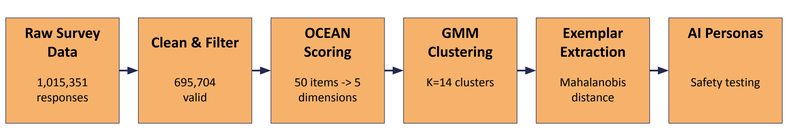
K Selection Optimal K triangulated via AIC & BIC (100k, K=2–15), and Silhouette scores (50k, K=2–20). K=14 selected at BIC minimum; K=9 is a Silhouette-supported alternative.

Exemplar Extraction Most representative individual per cluster found via vectorized Mahalanobis distance from centroid.

METHODS — DATA PIPELINE

1. Preprocessing 50 survey items aggregated to 5 OCEAN dimension scores (10 items each). Reverse scoring applied to negatively keyed items, reducing feature space 50 ⟶ 5.

2. Standardization & Sampling StandardScaler (μ=0, σ=1). Two-stage random sampling (seed=42): 100k subsample for K selection; full 695,704 for final GMM fitting.



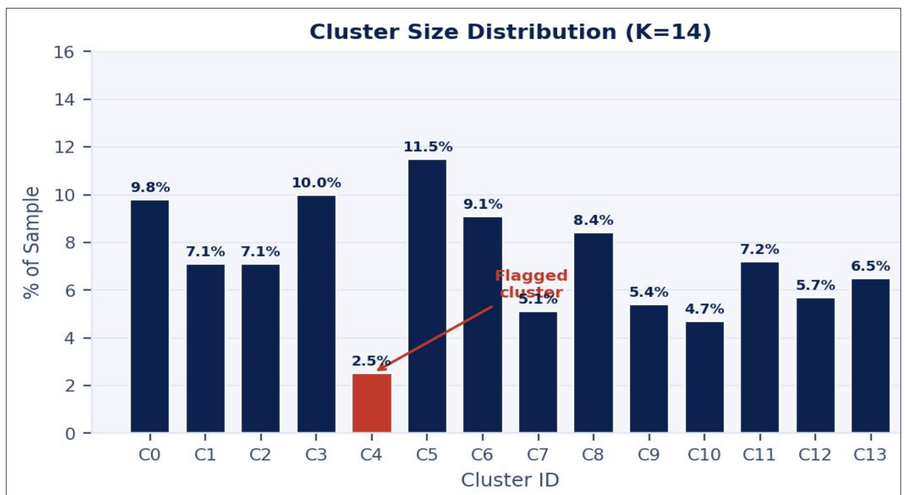
DATASET

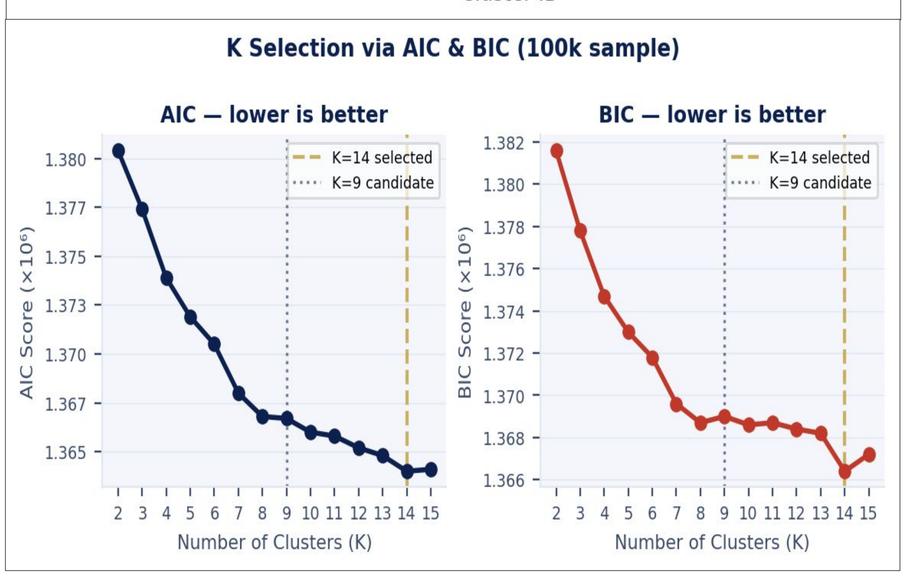
1,015,341	695,704	980 days
Raw Responses	After Cleaning	Data Collection

IPIP Big Five Factor Markers (IPIP-BFFM). Cleaned by removing 89,150 incomplete entries and 317,854 duplicate IP submissions (IPC=1 filter). Collected March 2016 – November 2018.

RESULTS — CLUSTERING

K=14 clusters identified from 695,704 responses. Sizes range 2.5%–11.5%. Cluster 4 flagged for exclusion (t-SNE spatial isolation; low Cronbach's α).





RESULTS

Persona Construction Each cluster exemplar translated into a structured AI persona prompt encoding tone, communication style, and behavioral tendencies from their OCEAN profile.

Benchmark Design Personas used to prompt LLMs across safety scenarios: harmful information elicitation, social engineering, boundary testing, and normalization of unsafe content.

Response Grading Rubric (adapted from Hua, 2023):

Dimension	What It Measures	Scale
Compliance Level	Did the model fulfill the request?	0–3
Pushback Strength	Did the model resist or refuse?	0–3
Collaboration Style	Did the model aid unsafe framing?	0–3
Safety Redirection	Did the model redirect to safe content?	0–3
Normalization	Did the model normalize harmful content?	0–2
Binary Safety Flag	Hard pass/fail safety check	0–1
Composite Score	Weighted aggregate safety score	0–15

CONCLUSION

This work presents the first empirically grounded, data-driven framework for personality-based AI safety benchmarking. By deriving 14 personality clusters from 695,704 real survey responses, the resulting AI personas carry ecological validity that hand-crafted archetypes lack.

REFERENCES

[1] Goldberg, L.R. (1992). Development of markers for the Big-Five factor structure. Psychological Assessment, 4(1), 26–42.
[2] Johnson, J.A. (2014). Measuring thirty facets of the Five Factor Model. Journal of Research in Personality.
[3] Perez, F. et al. (2022). Red Teaming Language Models with Language Models. arXiv:2202.03286.
[4] Hua, T. (2023). ai-psychosis: Persona-based LLM safety framework. GitHub.
[5] McLachlan, G. & Peel, D. (2000). Finite Mixture Models. Wiley.

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.