Commentary

Comment on Zhu J, Li X, Maharjan J, Coifman KG, Jin R. Benchmarking Personality Inference in Large Language Models Using Real-World Conversations. J Psychiatry Brain Sci. 2025;10(6):e250020

Chunyu Liu

Institute For Human Performance (IHP), Upstate Medical University, Syracuse, NY 13210, USA; liuch@upstate.edu

INTRODUCTION

This manuscript [1] presents a careful and timely benchmarking study of large language models (LLMs) for personality inference from real-world conversational data. The authors are commendable for reporting largely negative results transparently and for conducting extensive analyses across models, prompting strategies, and input lengths. However, several methodological gaps currently limit the interpretability of the findings.

Concerns

Validity of the Interview Paradigm for Personality Inference

While the semi-structured interviews are described as predictive of general psychological processes, the manuscript does not establish that they contain sufficient information for personality inference. No reference was provided for the design and validity test of the interview. Critically, no human baseline is provided. Without demonstrating that human raters can reliably infer personality traits from these transcripts, it is unclear whether weak LLM performance reflects model limitations or an ill-posed task. This distinction is fundamental for interpreting the benchmark.

Use of BFI-10 and Aggregated Big Five Traits as the Reference Standards

The BFI-10 is a highly abbreviated instrument with known reliability limitations. The manuscript does not discuss expected ceiling correlations, attenuation due to measurement error, or human-human agreement. As a result, reported correlations ($r \approx 0.2–0.3$) cannot be meaningfully interpreted as “poor” or “near ceiling.” Contextualizing model performance relative to reference reliability is essential. It is good to test correlation between BFI-10 and Big Five to set the expectation.

Open Access

Published: 4 Jan 2026

Item-Level Versus Trait-Level Agreement Discrepancy

Inter-model agreement is low at the BFI-10 item level but substantially higher at the aggregated Big Five level. This pattern raises concerns that aggregation may induce convergence toward shared normative or stylistic priors rather than valid trait inference. The manuscript should more explicitly address this construct-level implication.

Unresolved Effects of Input Length and Prompting

The finding that medium-length input outperforms full context, and that chain-of-thought prompting often degrades performance, remains insufficiently explained. Without hypothesis-driven ablations or alternative prompting controls, these effects remain descriptive rather than explanatory.

Unjustified Removal of Punctuation during Preprocessing

The decision to remove punctuation from transcripts prior to LLM inference is not justified and may be problematic. For LLMs, punctuation conveys syntactic, discourse, and affective information that is potentially relevant to personality expression. This preprocessing choice reflects legacy NLP practices rather than LLM-appropriate input handling and may inadvertently suppress informative signals. At minimum, the authors should justify this decision or provide an ablation comparing raw versus normalized transcripts.

CONFLICTS OF INTEREST

The author declares no conflicts of interest.

REFERENCES

1. Zhu J, Li X, Maharjan J, Coifman KG, Jin R. Benchmarking personality inference in large language models using real-world conversations. J Psychiatry Brain Sci. 2025;10(6):e250020.

How to cite this article:
Liu C. Comment on Zhu J, Li X, Maharjan J, Coifman KG, Jin R. Benchmarking personality inference in large language models using real-world conversations. J Psychiatry Brain Sci. 2025;10(6):e250020. J Psychiatry Brain Sci. 2026;11(1):e260001. https://doi.org/10.20900/jpbs.20260001.