LLMs Consistency in Describing and Scoring Personality

Danila Chernousov

April 2, 2025

Abstract

In this paper, we propose a novel approach for evaluating Large Language Models' ability to be profiled with a simple prompt containing a set of scores along several personality dimensions. We assess LLM consistency in both describing and scoring personality traits through a two-stage, cross-session experiment. We show that when asked explicitly, LLMs are able to convert numerical scores to text and back. Our framework evaluates LLM consistency not only on well-established traits (e.g., Big Five) but also on understudied or synthetic trait combinations, probing their robustness in novel psychological subspaces. We find that LLMs exhibit higher consistency for widely studied traits (e.g., Big Five) compared to under-researched or ambiguous constructs, suggesting a dependency on training data priors.

Keywords: AI Agents, Large Language Model, Agent-Based Simulation, Agent Profiling, LLM consistency.

1 Introduction

The ability to profile artificial agents with stable, interpretable personality traits represents a critical challenge in developing human-centered AI systems. Personality can be conceptualized as a point in a multidimensional vector space, providing a framework for agents to exhibit coherent behaviors and adapt interactions to user preferences. While psychological research has traditionally employed trait frameworks like the Big Five Inventory (openness, conscientiousness, extraversion, agreeableness, neuroticism) as basis vectors, this convention raises important questions when applied to modern Large Language Models (LLMs). These models are trained on vast corpora containing established psychological theories.

Existing research has predominantly evaluated LLM consistency through psychological questionnaires and established trait frameworks. While valuable, these approaches inherently tie assessments to human-centric constructs, potentially biasing evaluations toward traits with strong lexical or research priors. Our work introduces a novel methodology that decouples personality evaluation from pre-existing instruments, instead testing LLMs' *intrinsic* capacity to model traits as abstract, composable dimensions.

Our approach establishes personality profiling as a bidirectional task:

- **Trait-to-Text:** Generating textual descriptions that implicitly encode predefined personality scores

- **Text-to-Trait:** Reconstructing original scores from generated descriptions in isolation

This framework requires no questionnaires, labeled data, or alignment with legacy psychological models. It enables testing of arbitrary trait combinations—whether aligned with traditional frameworks or novel constructs (e.g., "stoicism" or "curiosity")—while avoiding assumptions about LLMs' internal representations. This flexibility is crucial given that LLMs' training data embeds diverse, often conflicting personality theories that may not conform to rigid psychological conventions.

Our framework's data-free nature makes it particularly valuable for auditing model robustness and identifying biases without requiring external validation. All of the code used can be found at the github page. [1]

2 Problem Statement

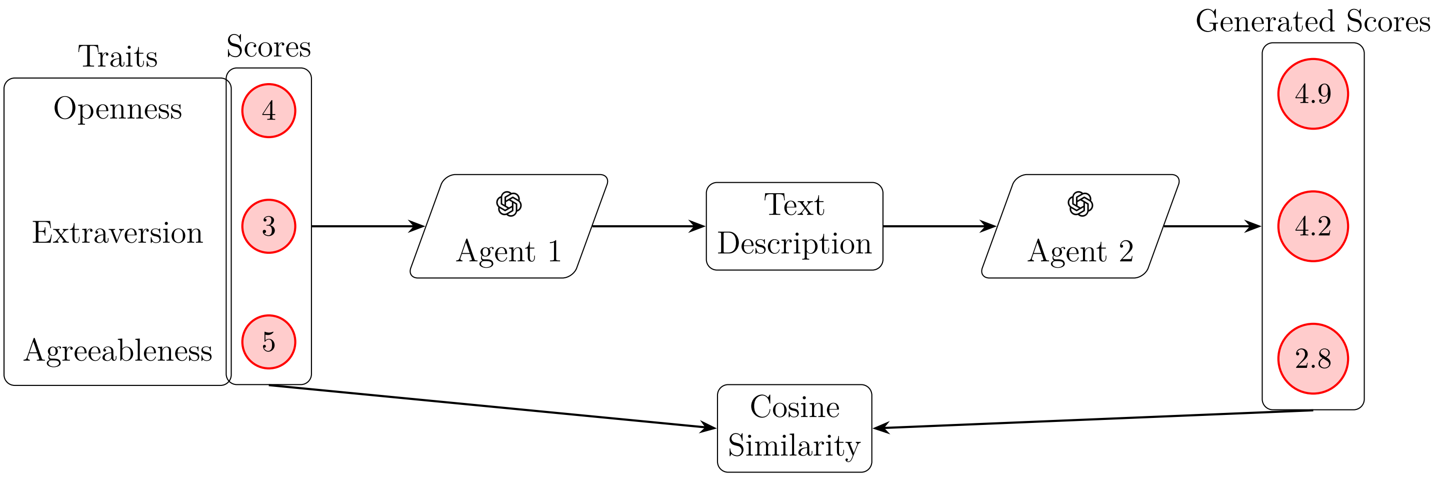

Figure 1: Experiment pipeline. We prompt Agent1 to generate text description. Agent2 generates scores based on the text description without seeing original scores

The problem can be formulated as follows:

Can LLMs be consistent in personality scoring?

To explain this idea further we will use mathematical notation

Let $\mathbf{p} \in \mathcal{P} \subset \mathbb{R}^n$ represent a personality trait vector in an n-dimensional space, and $\mathcal{D}$ denote the space of textual descriptions. We consider two mappings:

$$f : \mathcal{P} \to \mathcal{D}, \quad f(\mathbf{p}) = \mathbf{d} \quad \text{(trait-to-text encoding)} \tag{1}$$

$$g : \mathcal{D} \to \mathcal{P}, \quad g(\mathbf{d}) = \mathbf{p}' \quad \text{(text-to-trait decoding)} \tag{2}$$

[1] https://github.com/Chernousovdv/Personalized-AI-Agent

The core problem is to evaluate whether the composition $g \circ f$ preserves personality vectors with minimal distortion:

$$\|\mathbf{p} - g(f(\mathbf{p}))\| < \epsilon \quad \forall \mathbf{p} \in \mathcal{P}, \tag{3}$$

where ϵ is an acceptable error threshold. Specifically:

- Do LLMs implement f and g such that reconstructed vectors $\mathbf{p}'$ remain ϵ-close to originals $\mathbf{p}$?
- Does the error $\|\mathbf{p} - \mathbf{p}'\|$ increase for subspaces $\mathcal{S} \subset \mathcal{P}$ with limited training data coverage?

This quantifies whether LLMs can reliably serve as lossless translators between numeric personality representations and their textual embeddings - a fundamental requirement for stable AI agent design.

3 Theory

We use the following algorithm for each set of traits. The traits we chose are discussed in the next section

Algorithm 1: LLM Personality Consistency Evaluation

```
Input: Personality traits $\mathcal{T}$, True scores $S_{true}$, Number of trials $N$
Output: Consistency metrics $M$
for $i \leftarrow 1$ to $N$ do
    $D_i \leftarrow$ GenerateDescription($\mathcal{T}, S_{true}$)
    $S_{recon} \leftarrow$ ExtractScores($D_i, \mathcal{T}$)
    for $\tau \in \mathcal{T}$ do
        $M_i[\tau] \leftarrow |S_{true}[\tau] - S_{recon}[\tau]|$
    end
    UpdateMetrics($M, M_i$)
end
$M_{mae} \leftarrow$ MeanAbsoluteError($M$)
$M_{corr} \leftarrow$ CosineSimilarity($M$)
return $M_{mae}, M_{corr}$
```

4 Experiment

TODO experiment

5 Results

TODO Results

A Prompts

TODO prompts