

# LLMs Consistency in Describing and Scoring Personality

Danila Chernousov  
Expert: Yury Maximov  
Consultant: Ivan Novikov

Moscow Institute of Physics and Technology

2025

# Evaluating consistency

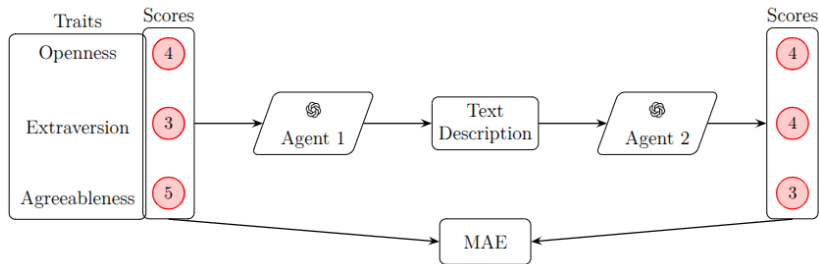
Develop a framework for measuring LLM fidelity in personality profiling for any set of personality traits.

Evaluate LLM consistency in converting numerical personality scores to text and back.

Assess LLM performance on both well-established (e.g., Big Five) and understudied/synthetic traits.

Investigate the impact of parameters (e.g., temperature) and trait ordering on consistency.

# Experiment pipeline






$S_{agent, traits} : \text{score} \rightarrow \text{text}$

$S_{agent, traits}^{-1} : \text{text} \rightarrow \text{score}$

1. Big Five – set of personality traits with a lot of research
2. Personality conditioning – prompting model with a personality

# Existing approaches

-  Bisbee, James et al. (May 2024). “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models”. In: *Political Analysis* 32, pp. 1–16. DOI: [10.1017/pan.2024.5](https://doi.org/10.1017/pan.2024.5).
-  Park, Joon Sung et al. (2024). “Generative agent simulations of 1,000 people”. In: *arXiv preprint arXiv:2411.10109*.
-  Serapio-García, Gregory et al. (2023). “Personality traits in large language models”. In.

# Evaluating LLM Consistency

## Core Challenges

1. LLMs lack intrinsic personality; their responses vary based on input structure (e.g., **trait order**, **prompt phrasing**).
2. Prompting agent with personality induces consistency.
3. No frameworks for evaluating LLMs along arbitrary personality dimensions

## Key Questions

1. Can LLMs reliably map between numerical scores ( $\mathbf{p}$ ) and textual descriptions ( $\mathbf{d}$ )?
2. Is the reconstruction error  $\|\mathbf{p} - g(f(\mathbf{p}))\| < \epsilon$  consistent across traits?
3. Does error increase for understudied traits ( $\mathcal{S} \subset \mathcal{P}$ )?

# Measuring Self-consistency

## Key advantages

1. Simplicity of the prompt: Less than 1000 tokens combined
2. Does not rely on learned associations
3. Can be applied to any set of traits without having to rely on questionnaires or datasets

# Temperature Experiment: Key Findings

## Experimental Setup

1. Varied temperature (0.2-1.5) during description generation
2. 200 runs with Big Five traits

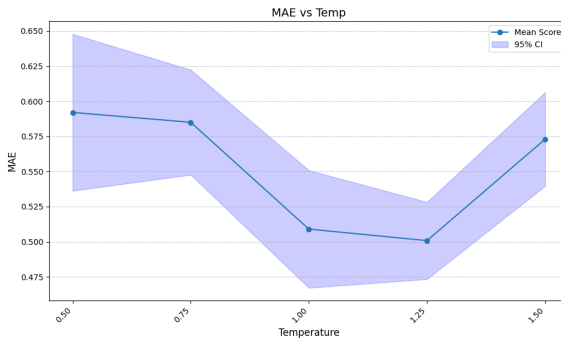
## Optimal Range Found

Best balance at **T = 1.25**

Preserves trait fidelity while allowing nuance

## Key Insight

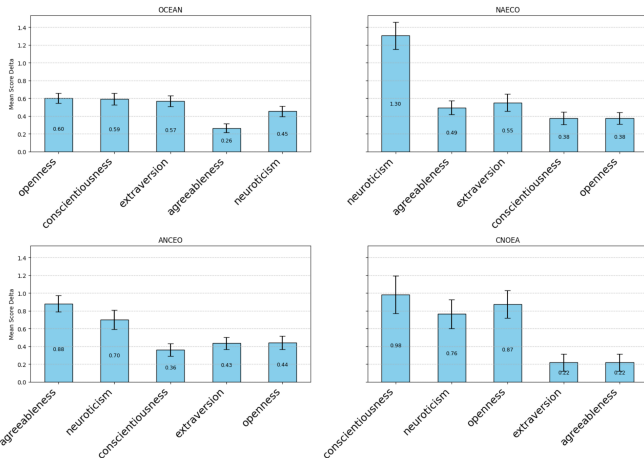
Temperature isn't just stylistic - it **directly affects** how much latent trait signal survives text generation.



MAE vs. Temperature

# Ordering Experiment: Impact of Trait Sequence

Tested 4 trait orderings with Big Five:

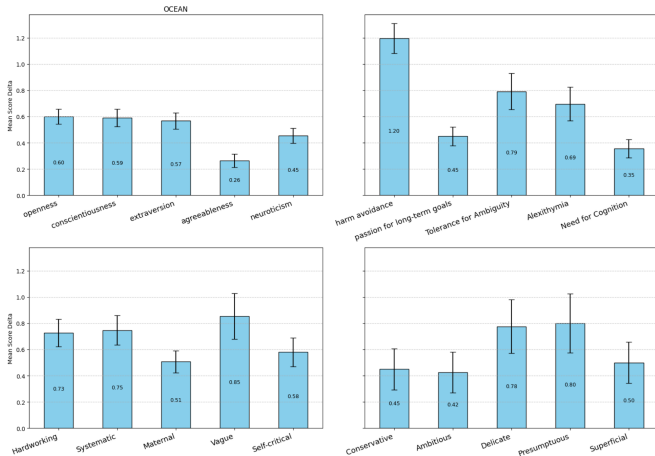


MAE across different trait orderings (O-C-E-A-N vs. alternatives).  
Key finding: Order effect exists and is significant.



# Consistency Across Different Trait Sets

Tested 3 other sets of traits



MAE across different trait categories (O-C-E-A-N vs. alternatives).  
Key finding: Model performs better on OCEAN.

# Conclusion & Limitations

## Key Findings

1. **Two-stage framework** effectively measures LLM consistency in personality simulation
2. **Significant ordering effects**
3. Impact of learned associations: Tested model performed best on traditional set of traits (OCEAN)

## Limitations

1. Single-model bias (DeepSeekV3 only)
2. Doesn't measure intermediate biases
3. Captures the "descriptiveness" trait – the degree to which it can be described without mentioning it directly