LLMs Consistency in Describing and Scoring Personality

Danila Chernousov

MIPT

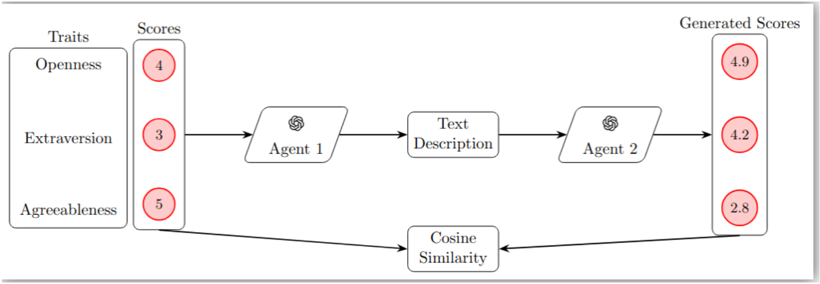
2025

Goals

- 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.

LLMs consistency in personality scoring

Experiment pipeline



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

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

Research question: $S(S^{-1}) = I$**?**

Literature

Couple of motivational sentences

The problem

to investigate ...

The method needs a proper name here

put the brief idea here

The solution

your results appears twice, as a promise here and as a contribution later

1) set ...,
2) put ...,
3) get ....

Evaluating LLM Consistency

Core Challenge

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

Key Questions

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

problem statement ends with quality criterion

Couple of motivational sentences

The problem

to investigate . . .

The method needs a proper name here

put the brief idea here

The solution

your results appears twice, as a promise here and as a contribution later

1) set . . . ,
2) put . . . ,
3) get . . . .

Temperature Experiment: Key Findings

Experimental Setup

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

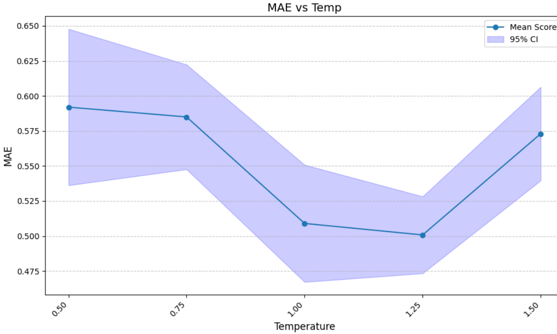


Figure: MAE vs. Temperature

Optimal Range Found

- Best balance at $\mathbf{T = 1.25}$
- Preserves trait fidelity while allowing nuance

Key Insight

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

Ordering Experiment: Impact of Trait Sequence

Experimental Design

- Tested 4 trait orderings with Big Five:

Key Finding

- Order effect exists and is significant

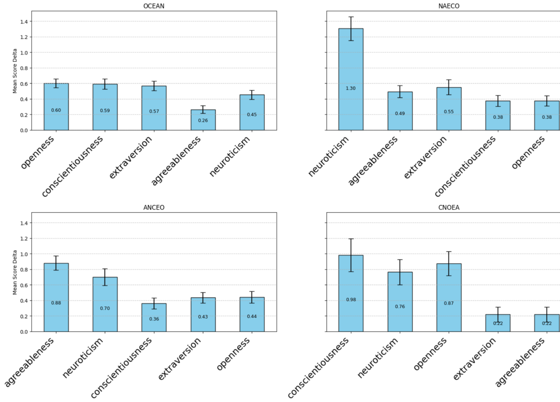


Figure: MAE across different trait orderings (O-C-E-A-N vs. alternatives)

Consistency Across Different Trait Sets

Key Findings

- Best performance on Big Five (MAE 0.26-0.6)
- Moderate success with clinical traits (MAE 0.4-1.2)
- Highest errors for synthetic traits (MAE up to 0.85)

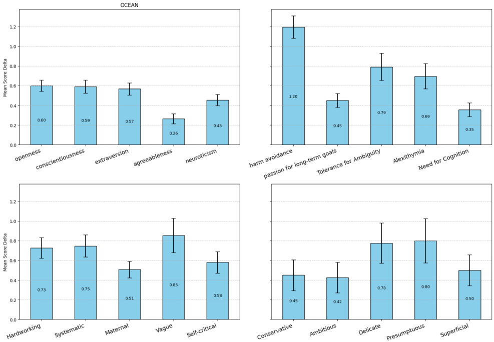


Figure: MAE across trait categories (your Fig.4 data)

Conclusion & Limitations

Key Findings

- **Two-stage framework** effectively measures LLM consistency in personality simulation
- Significant ordering effects: Up to 11% MAE variation across sequences
- Performance hierarchy:
  - Big Five (best) $\gg$ Clinical traits $\gg$ Abstract/Synthetic

Limitations

- Single-model bias (DeepSeekV3 only)
- Numerical MAE may mask semantic errors
- Ordering effects not tested on synthetic traits