LLMs Consistency in Describing and Scoring Personality

Danila Chernousov, Ivan Novikov

May 14, 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 experiment. We show that when asked explicitly, LLMs can convert numerical scores to text and back consistently. Our framework is able to evaluate LLM consistency not only on well-established traits (e.g., Big Five) but also on understudied or synthetic trait combinations.

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

1 Introduction

Large language models (LLMs) have been rapidly integrated into applications that require human-like interaction [Sun et al., 2024, Argyle et al., 2023, Park et al., 2023] fueling interest in their ability to simulate consistent, personalized behavior. Argyle et al. introduced the idea of silicone sampling, enabling LLMs to represent human populations in downstream tasks [Argyle et al., 2023, Santurkar et al., 2023, Manning et al., 2024, G. Jiang et al., 2023, Serapio-García et al., 2023], while [Horton, 2023, Leng, 2024, H. Jiang et al., 2023]. demonstrated that LLMs reflect many of the same biases observed in human decision-making.

If we imagine personality as a point in a vector space, we can choose basis vectors, such as openness, extraversion, and agreeableness and profile agents along these dimensions. In theory, this allows for precise control over agent behavior using a small set of trait values. However, before employing arbitrary or task-specific trait sets, we must be certain that an LLM both understands such profiling and behaves within the confines of the specified personality. A core challenge in agent profiling is configuring LLMs to adopt target personality traits, values, or behavioral patterns purely through prompts, without relying on long-term memory or fine-tuning.

Although recent work has explored the capacity of LLMs to mimic human personalities, a critical question remains understudied: *consistency* — the degree to which an agent simulates the behavior of a personality with which it was profiled. Existing research [Frisch and Giulianelli, 2024, Tommaso et al., n.d., Shu et al., 2023, Song et al., 2023 H. Jiang et al., 2023, Gupta et al., 2023] indicates that LLMs do not possess any personality by themselves and thus are not consistent (for example, their answers depend on the order in which they were presented to them). Yet, existing evaluations have largely focused on alignment with established psychometric frameworks like the Big Five Inventory (BFI) - most popular psychometric scale, consisting of 5 traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). These approaches share two key limitations:

Memorized Associations: They rely on well-documented traits—e.g., extraversion or neuroticism—for which LLMs have abundant training data, potentially conflating true consistency with learned textual correlations.

Limited Scope: Restricting analysis to predefined human-centric taxonomies precludes assessment of synthetic or understudied trait combinations (e.g., "risk-seeking vs. deliberative" or "aesthetic sensitivity"), leaving gaps in our understanding of LLMs' general profiling capabilities.

To address these gaps, we introduce a two-stage, cross-session methodology that evaluates LLMs' intrinsic consistency in translating between numerical personality scores and textual descriptions—and back again—without relying on contextual memory. In the first stage, the model receives a prompt containing specific trait scores and generates an implicit description of an agent's personality. In the second stage, the same model is tasked with inferring numerical scores based solely on that description. Each stage occurs in a separate chat session, ensuring no hidden memory of previous inputs. By comparing original and inferred scores, we directly measure the model's consistency in both describing and interpreting personality traits.

This framework functions as an intrinsic benchmark of an LLM's ability to convert structured personality information (i.e., numerical scores) into a coherent, text-based persona, and to reverse this process when prompted. It also allows us to test the limits of LLM flexibility across both well-established and non-canonical dimensions.

Applying our framework to both canonical (e.g., Big Five) and novel trait sets, we find that LLMs exhibit higher consistency for widely studied traits. Additionaly, we show that consistency depends on the ordering of traits. These results highlight a tension between LLMs' flexible role-playing capabilities and their dependence on patterns embedded in training data.

All of the code used can be found at the github page. [1]

2 Problem Statement

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:

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

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

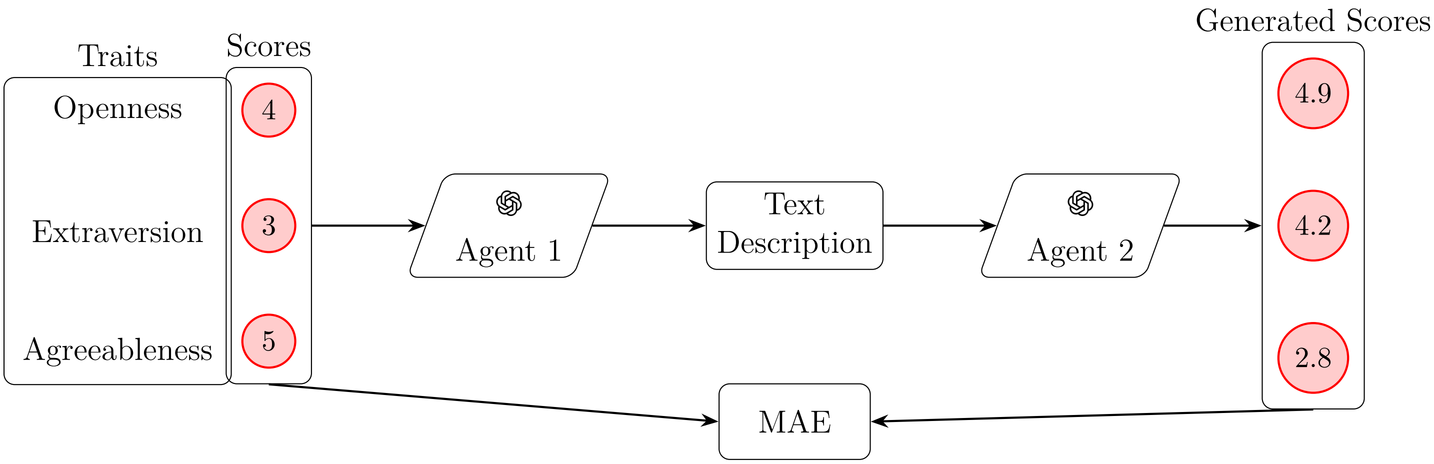

Figure 1: Experiment pipeline. We prompt Agent1 to generate text description. Agent2 generates scores based on the text description without seeing original scores

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

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($M$)
return $M_{mae}$
```

4 Role of Temperature in Generating Descriptions

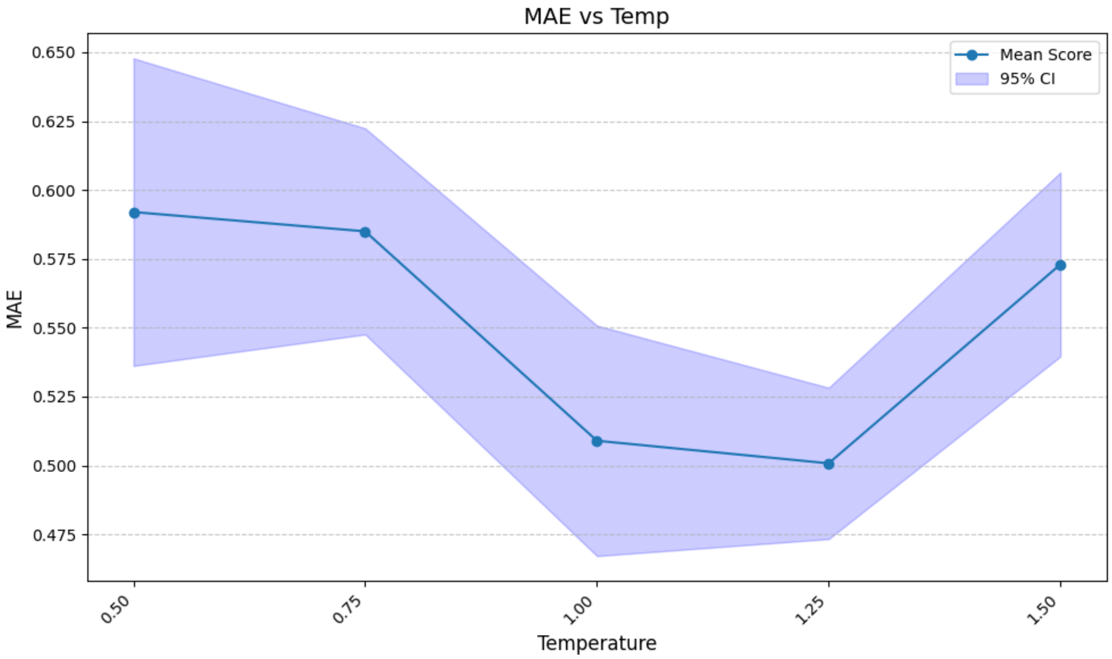

Figure 2: Mean Average Error on Big-Five traits set vs. Temperature parameter used for description generation. It c

An important factor influencing the behavior of large language models during text generation is the *temperature* parameter. Temperature controls the degree of randomness in the model's output by scaling the probabilities of the next token in the sequence. Lower temperatures (e.g., 0.2–0.5) make the model more deterministic and focused, often favoring high-probability completions. In contrast, higher temperatures (e.g., 0.8–1.2) increase diversity by allowing less likely tokens to be sampled more frequently, resulting in more varied and creative but also less predictable outputs.

In the context of our framework, temperature plays a crucial role during the first stage, where the model generates a natural language *description* based on a given set of personality trait scores. Since these descriptions serve as the input for subsequent trait inference, their fidelity to the original traits directly impacts the overall consistency score. A high temperature may lead to imaginative or stylistically rich descriptions that drift from the intended personality profile, while a low temperature may result in repetitive or overly rigid phrasing that underrepresents certain nuances of the personality.

To quantify the impact of temperature on descriptive consistency, we conducted several experiments using the Big Five personality traits. Each experiment varied the temperature parameter during the description generation phase, and consisted of 200 runs to ensure statistical significance. We compared outcomes using Mean Absolute Error (MAE) between the original input scores and the scores inferred from the generated descriptions.

Our results indicate that intermediate temperatures (around 1.25) strike the best balance between expressiveness and trait fidelity. At this setting, descriptions are rich enough to reflect subtle trait cues while remaining grounded in the original score profile. Consequently, we adopted this temperature value for all subsequent experiments to ensure consistent and comparable evaluation across canonical and synthetic trait sets.

This result underscores an important insight: fidelity in reverse inference is not solely a function of linguistic precision or predictability. Rather, it depends on whether the generated text encodes enough semantically rich cues for the model to reconstruct the original intent. Temperature, in this sense, is not just a stylistic parameter, it plays a direct role in shaping how much of the latent trait signal is preserved in the text.

5 Order Variation

In our second experiment, we investigate the effect of trait ordering on LLM consistency, inspired by the findings of [Gupta et al., 2023], who observed that LLM responses may vary depending on the sequence in which information is presented. To test this within our framework, we generate descriptions from the same set of Big Five personality scores, but vary the order in which traits are presented in the input prompt.

We evaluate four different orderings, including the canonical O-C-E-A-N sequence, as well as three alternative permutations. The inference step remains identical across conditions, using the generated description (from each ordering) to recover the original scores.

Our results show that trait order introduces major challenge when working with LLMs as they display different behaviour when presented with shuffled prompts.

6 Different Personality Dimensions

In this experiment, we extend our evaluation beyond canonical personality constructs to include a variety of lesser-known or synthetic dimensions. While the Big Five model (OCEAN) serves as a well-established benchmark, real-world deployments often require LLMs to reason

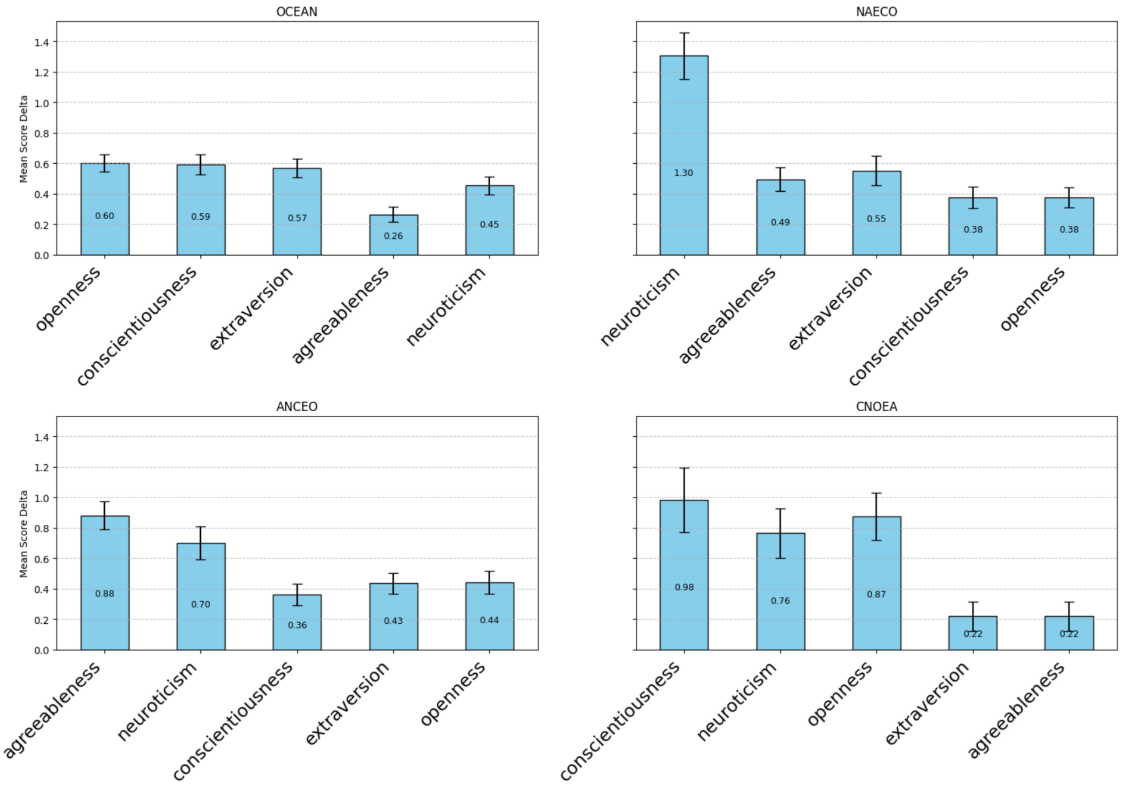

Figure 3: Mean Average Errors on Big-Five with different ordering of traits. In the top left corner is the most popular choice - OCEAN

about personality traits that lack widespread theoretical grounding or consistent representation in training data. This experiment tests the adaptability and generalization of LLMs across such dimensions.

Figure4 presents the mean score deltas (original vs. inferred) for four different sets of traits. The top-left plot (OCEAN) provides a baseline, with most traits reconstructed with moderate accuracy (errors around 0.5–0.6), and agreeableness showing the highest fidelity (0.26). In contrast, the three alternative sets show more varied and often higher error rates.

The top-right plot includes constructs like *harm avoidance* and *alexithymia*, which are less commonly encountered in general discourse. Here, *harm avoidance* exhibits a pronounced error (1.20), suggesting significant difficulty in encoding this trait from a textual description. Conversely, *passion for long-term goals* and *need for cognition* are inferred with relatively low error.

The bottom-left plot consists of more abstract or ambiguous constructs like *vague*, *maternal*, and *self-critical.* Traits such as *vague* result in higher errors (0.85), indicating interpretability challenges.

In the bottom-right plot, we present another synthetic trait set. Errors here remain moderate to high (up to 0.80), and standard deviations are often large, reflecting inconsistent reconstructions. This aligns with our broader observation that the less canonically defined or semantically diffuse a trait is, the more prone the LLM is to either misrepresent or default to stereotypical interpretations.

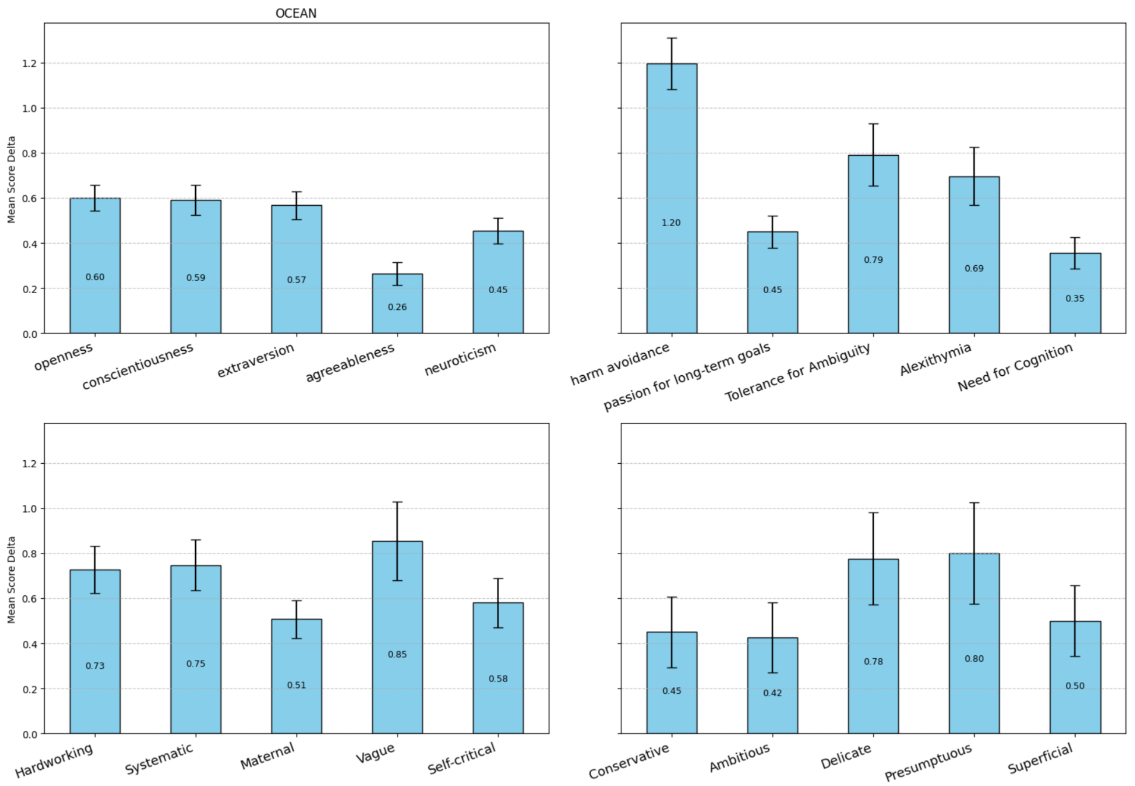

Figure 4: Mean Average Errors on Big-Five with different ordering of traits. In the top left corner is the most popular choice - OCEAN

These results emphasize a key limitation: while LLMs can encode and reconstruct widely known personality traits with reasonable consistency, their generalization to non-canonical or synthetic dimensions is less reliable.

7 Conclusion

In this study, we introduced a novel two-stage framework to evaluate the consistency of Large Language Models (LLMs) in describing and scoring personality traits. By decoupling the processes of trait-to-text encoding and text-to-trait decoding across independent sessions, our methodology provides a robust, intrinsic benchmark for assessing LLMs' ability to simulate stable and coherent personalities. This framework advances the field by addressing key limitations of prior work, such as reliance on memorized associations and the narrow scope of predefined human-centric taxonomies.

Our experiments revealed that LLMs exhibit higher consistency for well-established traits like the Big Five, while struggling with ambiguous or synthetic constructs. This underscores the tension between their flexible role-playing capabilities and their dependence on patterns embedded in training data. Additionally, we demonstrated the critical role of the temperature parameter in balancing expressive richness and trait fidelity, with intermediate values (e.g., 1.25) yielding the most reliable results. Trait ordering, had a significant impact on consistency.

The development of this framework opens new avenues for research and application. It enables precise control over AI agent personalities, facilitating their use in simulations, virtual assistants, and other interactive systems where behavioral consistency is paramount. Future work could explore extensions to dynamic or context-dependent personality modeling, as well as the integration of multimodal inputs for richer trait representations.

Ultimately, our findings highlight both the promise and limitations of LLMs in personality simulation, providing a foundation for more reliable and transparent AI agent design. The framework's adaptability to diverse trait sets positions it as a valuable tool for advancing the study of artificial personalities and their real-world applications.

8 Limitations

First, our experiments were conducted exclusively using the DeepSeekV3 API. While this allowed for controlled and reproducible results, it limits the generalizability of our findings. Different LLMs (e.g., GPT-4, Claude, Gemini) may exhibit varying degrees of consistency due to differences in architecture, training data, and alignment techniques. Future work should expand evaluations to multiple models to assess whether our framework yields similar results across different systems.

Additionally, our consistency metric relies on numerical reconstruction error, which may not fully capture semantic fidelity. A low error score does not necessarily mean the generated descriptions are contextually appropriate or nuanced. Human evaluations could complement automated metrics to assess qualitative aspects of personality simulation.

Addressing these limitations in future work will strengthen the robustness and applicability of our framework, enabling more reliable profiling of LLM-based agents across diverse use cases.

A Prompts

Description prompt
Analyze the following personality description and provide numerical scores between 1 (low) and 5 (high) for each of these traits: <traits_str>.

Description: <description>

Format your response as a JSON dictionary with the trait names as keys (use lowercase) and the scores as values. Do not include any additional text or explanations. Example response format: <example_json>

Scoring prompt
Create a detailed personality description matching these exact trait scores: <traits_with_scores>

Express these traits however you like but you MUST NOT mention these scores as numbers explicitly. Your goal is to describe these scores in words in a best possible way Critical constraints: - Never reference numbers or scores in the description - Avoid direct trait mentions (e.g., don't say 'extraverted') - Show personality through specific, contextual details - Maintain psychological plausibility - Include nuanced contradictions when appropriate

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Frisch, I., & Giulianelli, M. (2024). Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

Gupta, A., Song, X., & Anumanchipalli, G. (2023). Self-assessment tests are unreliable measures of llm personality. *arXiv preprint arXiv:2309.08163*.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Tech. rep.). National Bureau of Economic Research.

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2023). Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, *36*, 10622–10643.

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2023). Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Leng, Y. (2024). Can llms mimic human-like mental accounting and behavioral biases? *Available at SSRN 4705130*.

Manning, B. S., Zhu, K., & Horton, J. J. (2024). *Automated social science: Language models as scientist and subjects* (tech. rep.). National Bureau of Economic Research.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *International Conference on Machine Learning*, 29971–30004.

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models.

Shu, B., Zhang, L., Choi, M., Dunagan, L., Logeswaran, L., Lee, M., Card, D., & Jurgens, D. (2023). You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *arXiv preprint arXiv:2311.09718*.

Song, X., Gupta, A., Mohebbizadeh, K., Hu, S., & Singh, A. (2023). Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms. *arXiv preprint arXiv:2305.14693*.

Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., & Kim, J. H. (2024). Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*.

Tommaso, T., Hegazy, M., Lemay, D., Abukalam, M., Rish, I., & Dumas, G. (n.d.). Llms and personalities: Inconsistencies across scales. *NeurIPS 2024 Workshop on Behavioral Machine Learning*.