

CAPE: Context-Aware Personality Evaluation Framework for Large Language Models

Jivnesh Sandhan, Fei Cheng, Tushar Sandhan[†] and Yugo Murawaki

Kyoto University, Japan and [†]IIT Kanpur, India

{jivnesh, feicheng, murawaki}@i.kyoto-u.ac.jp, [†]sandhan@iitk.ac.in

Abstract

Psychometric tests, traditionally used to assess humans, are now being applied to Large Language Models (LLMs) to evaluate their behavioral traits. However, existing studies follow a context-free approach, answering each question in isolation to avoid contextual influence. We term this the Disney World test, an artificial setting that ignores real-world applications, where conversational history shapes responses.

To bridge this gap, we propose the first Context-Aware Personality Evaluation (CAPE) framework for LLMs, incorporating prior conversational interactions. To thoroughly analyze the influence of context, we introduce novel metrics to quantify the consistency of LLM responses, a fundamental trait in human behavior.

Our exhaustive experiments on 7 LLMs reveal that conversational history enhances response consistency via in-context learning but also induces personality shifts, with GPT-3.5-Turbo and GPT-4-Turbo exhibiting extreme deviations. While GPT models are robust to question ordering, Gemini-1.5-Flash and Llama-8B display significant sensitivity. Moreover, GPT models response stem from their intrinsic personality traits as well as prior interactions, whereas Gemini-1.5-Flash and Llama-8B heavily depend on prior interactions. Finally, applying our framework to Role Playing Agents (RPAs) shows context-dependent personality shifts improve response consistency and better align with human judgments.¹

1 Introduction

Large Language Models (LLMs) have made significant advances in generating human-like text (Spangher et al., 2024; Ou et al., 2024). Moving beyond linguistic fluency, this raises the fundamental question: “How human-like are LLMs?” Researchers are utilizing psychometrics to assess

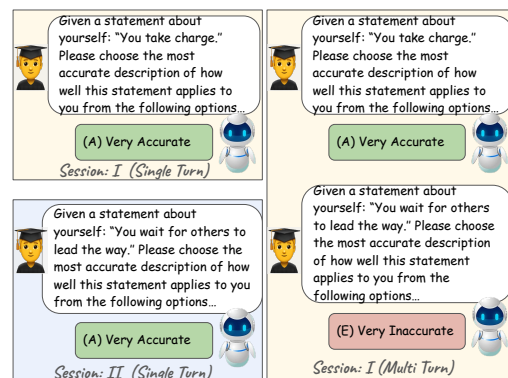


Figure 1: Illustration of the difference between existing context-free (left) and the proposed context-aware (right) evaluation framework. In the left setting, each question is asked independently, with no prior history (indicated by a different background color). In the right setting, all questions are part of the same session, where prior answers can influence future responses.

whether LLMs exhibit human-like personality traits (Huang et al., 2024b; Serapio-García et al., 2023). This question is crucial in contexts where LLMs may act as human proxies in surveys (Dillion et al., 2023; Harding et al., 2024) or personalized human-AI interactions (Tseng et al., 2024).

Psychology provides well-established personality assessment frameworks, such as the Big Five Personality model (McCrae and John, 1992), which evaluate individuals through a series of questions rated on a Likert scale. Analogously, LLMs are assessed using a zero-shot multiple-choice question (MCQ) format (§2). However, the reliability of these assessments remains a subject of debate. Some researchers advocate for methodologies that measure intrinsic personality traits in LLMs (Jiang et al., 2023; Wang et al., 2024b; Yang et al., 2023; Jiang et al., 2024), while others highlight inconsistencies stemming from prompt sensitivity (Shu et al., 2024; Gupta et al., 2024; Song et al., 2023).

We identify a critical research gap in existing

¹Our code and datasets are publicly available at: <https://github.com/jivnesh/CAPE>

research: LLM personality assessments are conducted in an isolated, context-independent manner. We term this the *Disney World* test setting, where each question is answered without influence from prior responses. In contrast, real-world applications of LLMs necessitate exposure to conversational history. Before deploying LLMs in critical domains such as education and healthcare, it’s vital to understand how conversational history impacts personality assessments.

To bridge this gap, we propose the first Context-Aware Personality Evaluation (CAPE) framework, where prior questions and responses are retained in the conversational history to evaluate their impact on LLM personality (§3.1). To thoroughly analyze the influence of context, we assess response consistency, a fundamental human trait, by introducing various inconsistency factors related to prompt sensitivity such as temperature, option wording, option order, instruction and item paraphrasing. To quantify this consistency, we present novel metrics that measure the response pattern similarity across multiple LLM runs (§3.2).

Our key findings on the impact of context on LLM personality are: Prior conversational history enhances response consistency by serving as few-shot in-context learning (§6.1). However, introducing context leads to deviations in responses compared to context-independent setting, with GPT-3.5-Turbo and GPT-4-Turbo exhibiting extreme personality shifts (§6.2). Additionally, GPT models maintain intrinsic personality despite contextual influence, whereas Gemini-1.5-Flash and Llama-3.1-8B rely heavily on prior conversation (§6.3). While these GPT models remain robust to question ordering, Gemini-1.5-Flash and Llama-3.1-8B display significant sensitivity (§6.4). Finally, we demonstrate our framework on Role Playing Agents (RPAs) and show context-dependent personality shifts improve response consistency and better align with human scores (§6.5). Our key contributions are:

- To the best of our knowledge, we introduce the first Context-Aware Personality Evaluation (CAPE) framework, demonstrating its role in enhancing consistency (§3).
- We propose novel metrics to quantify the consistency of LLM in assessments (§3.2).
- We conduct an in-depth analysis of how context influences LLM’s personality (§6).

2 Preliminaries: LLM’s Personality Test

The Big Five personality framework (McCrae and John, 1992; John and Srivastava, 1999) characterizes human personality using 5 fundamental traits: Openness (artistic, imaginative), Conscientiousness (organized, thorough), Extraversion (assertive, talkative), Agreeableness (appreciative, kind), and Neuroticism (anxious, worrying), collectively known as OCEAN. Following earlier works (Huang et al., 2024a; Jiang et al., 2023; Zhou et al., 2023), we assess the personality of an LLM by formulating the evaluation as a zero-shot multiple-choice question-answering task. Each assessment item consists of a self-descriptive statement and a set of response options. The model is prompted to evaluate how accurately the statement aligns with its personality by selecting the most appropriate response. The prompt for it as:

Given a statement about yourself: “You {Item}.” Please select the most accurate description of how well this statement applies to you from these options:

- (A) *Very Accurate*
- (B) *Moderately Accurate*
- (C) *Neither Accurate Nor Inaccurate*
- (D) *Moderately Inaccurate*
- (E) *Very Inaccurate*

where the Item describes behavioral tendency from a second-person perspective. Each item corresponds to one of the 5 OCEAN dimensions and is either positively (+Key) or negatively (-Key) related to that dimension. For example, (-O): “Do not like poetry” and (+O): “Love to daydream” are negatively and positively correlated to openness, respectively.

The responses are numerically scored based on their alignment with the corresponding trait dimension. If an item is positively correlated, options (A) to (E) are scored from 5 to 1; otherwise, if negatively correlated, they are scored from 1 to 5. For an assessment consisting of m items, the sequence of scores assigned to an LLM’s responses forms a **scoring trajectories**, which serve as the foundation for consistency analysis in our proposed framework. Mathematically, we define the scoring trajectory as, $\mathcal{T} = [s_1, s_2, \dots, s_m]$ where $\mathcal{T}$ is the scoring trajectory of the LLM, s_i is the score assigned to the LLM’s response for item i and m is the total number of items. This trajectory captures

the model’s response pattern across all items and is later utilized in our framework to analyze the consistency of LLM-generated personality.

For a given OCEAN trait d (where $d \in \{O, C, E, A, N\}$), the trait score of an LLM is computed as the average score across all items associated with that dimension: $S_d = \frac{1}{N_d} \sum_{i=1}^{N_d} s_i$, where S_d is the OCEAN score for trait d , N_d is the total number of items related to trait d , s_i is the score assigned to the LLM’s response for item i . The final **OCEAN score** of an LLM consists of the 5 computed scores (S_O, S_C, S_E, S_A, S_N), which provide a quantitative representation of the model’s personality tendencies. We also consider all permutations of OCEAN to make its associated trajectory order invariant (§3.2). These trajectories are then analyzed for consistency and alignment.

3 The Proposed Framework: Context-Aware Personality Evaluation

In our proposed framework, we introduce a Context-Aware Personality Evaluation (CAPE), where prior questions and responses are retained in the conversational history. To analyze the impact of context, we evaluate response consistency, a fundamental human trait, by considering various inconsistency factors related to prompt sensitivity, such as temperature, option wording, option order, instructions, and item paraphrasing. Finally, we introduce novel metrics to quantify this response consistency by measuring the similarity of response patterns across multiple LLM runs. We recognize that “context” can indeed encompass various dimensions, including interlocutor attributes or external databases, as indicated in broader NLP research. However, our study specifically focuses on the influence of conversational history on personality assessments in LLMs, since our primary objective is to capture the influence of conversational history.

3.1 Context-Aware Evaluation

Traditional LLM personality assessments treat responses in isolation, ignoring prior interactions. However, real-world applications involve multi-turn conversations where the context shapes responses. To bridge this gap, we propose a context-dependent personality assessment framework that retains prior exchanges while answering new questions (Figure 1). This approach enables a more *realistic* evaluation by assessing whether LLMs maintain consistent personality traits or shift based on

context—critical for applications like AI tutoring, virtual assistants, and social chatbots. Formally, let $Q = \{q_1, q_2, \dots, q_m\}$ be a set of personality assessment questions and $R = \{r_1, r_2, \dots, r_m\}$ the corresponding LLM responses. At any time step t , the conversational history is defined as $H_{t-1} = \{(q_1, r_1), (q_2, r_2), \dots, (q_{t-1}, r_{t-1})\}$, and the LLM’s response function becomes $r_t = f(q_t, H_{t-1})$, incorporating prior interactions.

Inconsistency Factors We introduce 5 sensitivity factors: temperature, option wording, option order, instruction, and item paraphrasing—each with 3 variants. For each variant, we generate 3 scoring trajectories using the same question order across independent LLM runs. We assess response consistency with multiple metrics, with temperature set to 0 except when testing its sensitivity. Refer to Appendix D for detailed prompts and examples.

(1) Stability: We establish a baseline by running the assessment 3 times without sensitivity factors, providing a reference for their impact.

(2) Temperature: We test 3 temperature values: 0.5, 1, and 1.5. While prior studies commonly use a default temperature of 0, real-world applications may require non-zero temperatures for more dynamic and adaptable behavior (Lee et al., 2025).

(3) Option Wording: We experiment with 3 paraphrased versions of each option while maintaining semantic equivalence (Shu et al., 2024). For example, “*Strongly agree*” is reworded as “*Completely Aligned*” and “*Perfectly Compatible*.”

(4) Option Order: We explore three ordering variations: original order ($A B C D E$), reverse order ($E D C B A$), and a randomized order ($C B D A E$) (Gupta et al., 2024; Song et al., 2023).

(5) Instruction: Previous studies have used different instruction formulations for personality evaluation (Huang et al., 2024a; Jiang et al., 2023; Serapio-García et al., 2023). We assess consistency across 3 variations of the instruction prompt.

(6) Item Paraphrasing: Using GPT4, we generate 2 paraphrased versions of each item (Huang et al., 2024a). For example, the original item “*Worry about things*” is reworded as “*Have anxiety about situations*” and “*Stress over issues*.” We manually verify paraphrased items for semantic fidelity to preserve its validity and reliability.

3.2 The Proposed Consistency Metrics

Standard consistency metrics that measure exact response agreement (Atil et al., 2024) or Eu-

clidean distance between trajectories overlook partial agreement and contextual dependencies between scores, treating all divergences equally. In response, we propose 2 novel metrics: Trajectory Consistency (TC) and OCEAN Consistency (OC). Our metrics focus on capturing the similarity in patterns of responses across multiple runs. By applying Gaussian Process Regression (GPR) independently on each trajectory, we account for both the responses and their interactions with neighboring questions. Our metrics assess consistency by evaluating the ratio of the intersection to the union of the posterior predictive distribution’s support at each point. A higher consistency corresponds to a greater overlap (intersection) of confidence intervals, while increased inconsistency results in a larger union (wider spread) of these intervals. This proportional relationship provides a clear measure of how consistent the model’s responses are across different assessment runs.

Trajectory Consistency (TC): Trajectory consistency refers to the similarity between three scoring trajectories produced when the same LLM takes the same psychometric test three times. Inconsistencies often arise due to inherent stochasticity in LLM outputs, sensitivity to prompt variations (e.g., option wording, order, or instructions), and lack of contextual grounding—all of which can cause response shifts despite an unchanged assessment. This is problematic because, like humans, a stable personality should yield consistent answers across repeated assessments; fluctuations undermine the reliability and interpretability of model behavior. Each LLM is assessed three times, producing three scoring trajectories. A trajectory is considered more consistent when these are closer in pattern and distance, which we quantify using our proposed Trajectory Consistency metric. A higher score indicates greater response stability.

Each scoring trajectory $\mathcal{T}_i$ represents an independent run of the LLM’s personality assessment, where $i \in \{1, 2, 3\}$. Each trajectory consists of pairs $(x_t, y_{i,t})$ for $t \in \{1, \dots, m\}$, where x_t is the index of t -th question, and $y_{i,t}$ is the score assigned in the i -th run. We apply a moving average filter with a window size ω ² for outlier denoising: $\hat{y}_{i,t} = \frac{1}{\omega} \sum_{j=0}^{\omega-1} y_{i,t-j}$, followed by mean normalization: $\hat{y}_{i,t} = \frac{\hat{y}_{i,t} - \mu_i}{\sigma_i}$ where μ_i and σ_i are the mean and standard deviation of the corresponding smoothed

trajectory. Then, we model each normalized trajectory using GPR as $f_i(x) \sim \mathcal{GP}(\mu_i(x), k_i(x, x'))$, where $\mu_i(x), k_i(x, x')$ are mean and kernel function.³ It gives the posterior predictive distribution at each x_t as $f_i(x_t) \sim \mathcal{N}(\mu_i(x_t), \sigma_i^2(x_t))$, where $\mu_i(x_t), \sigma_i^2(x_t)$ are posterior mean and variance respectively. We define *support interval* as $S_i(x_t) = [L_i(x_t), U_i(x_t)]$.⁴ The intersection of all 3 supports at x_t is

$$W_{\text{int}}(x_t) = \max\left(0, \min_i U_i(x_t) - \max_i L_i(x_t)\right) \quad (1)$$

The union of all supports is computed by first sorting the support intervals in ascending order based on their lower bounds and then iteratively merging the overlapping intervals followed by summing the lengths of all merged, non-overlapping supports as

$$W_{\text{union}}(x_t) = \sum_j \left(U_j^{\text{merged}}(x_t) - L_j^{\text{merged}}(x_t) \right) \quad (2)$$

where $U_j^{\text{merged}}(x_t)$ and $L_j^{\text{merged}}(x_t)$ represent the upper and lower bounds of the merged segments. Finally, consistency score TC is calculated as

$$TC = \frac{1}{x_m} \int_0^{x_m} \frac{W_{\text{int}}(x_t)}{W_{\text{union}}(x_t)} dx. \quad (3)$$

OCEAN Consistency (OC): We obtain an OCEAN score from each scoring trajectory (§2) and write it as: $\mathbf{s}_i = (O_i, C_i, E_i, A_i, N_i)$, $i \in \{1, 2, 3\}$. We generate all possible orderings of the 5 traits to make this representation order-invariant as $\mathcal{P}(\mathbf{s}_i) = \{\pi_j(\mathbf{s}_i)\}_{j=1}^5! = \{\mathbf{s}_{i,j}\}_{j=1}^{120}$. Then, we build a time series by appending each permuted sequence $\mathbf{s}_{i,j}$ for the i -th trajectory as: $\mathcal{T}_i = \{(x_t, y_{i,j,t})\}_{t=1}^{5 \times 120}$, where $y_{i,j,t}$ represents the score at position t in the permuted sequence $\mathbf{s}_{i,j}$. This transformation makes the series representation order-invariant. Then, we plug this series in the above formulation to obtain the consistency score. We call this as OCEAN consistency score.

4 Experimental Setup

Datasets: We use 2 datasets: the Machine Personality Inventory (MPI) (Jiang et al., 2023) (§5), licensed under MIT, which includes 120 items from the International Personality Item Pool (IPIP) and its IPIP-NEO adaptations (Goldberg, 1999; McCrae and Costa, 1997), and the Big Five Inventory (BFI) (§6.5) with 44 items (Lang et al., 2011).

³We use the Radial Basis Function kernel and automate hyper-parameter tuning with the scikit-learn library.

⁴For example, for the 95% confidence region, $L_i(x_t) = \mu_i(x_t) - 1.96\sigma_i(x_t)$, $U_i(x_t) = \mu_i(x_t) + 1.96\sigma_i(x_t)$.

²We use $\omega = 4$ based on our hyper-parameter tuning.

Construct validity and reliability of the psychometric instruments used: We would like to clarify that our work does not introduce new psychometric instruments, but rather builds on top of well-established ones in a more realistic, context-dependent evaluation setting. The validity and reliability of these instruments—such as the IPIP and BFI—have already been demonstrated in prior work on LLMs under context-independent settings. For example, Serapio-García et al. (2023) conducted a large-scale study across 18 LLMs showing strong construct validity and reliability of psychometric assessments. Similarly, ? reported high convergent validity between LLM-based and human-reported personality scores. Moreover, a growing body of research (Jiang et al., 2023; Wang et al., 2024b; Yang et al., 2023; Jiang et al., 2024) has consistently applied these psychometric instruments to LLMs, further reinforcing their validity and reliability. Our framework builds directly on these validated instruments, differing only in that it adopts a human-like interaction setting where conversational history is retained.

Systems: To evaluate LLM personality and enhance the generalizability of our findings, we select 7 diverse LLMs that vary in architecture, alignment strategies, and model size:: GPT-3.5-Turbo (OpenAI, 2022), GPT-4-Turbo (OpenAI, 2024), Gemini-1.5-Flash (Team et al., 2024), Claude-3.5-Haiku (Anthropic, 2024) and LLaMA-3.1-8B, Llama-3.3-70B, Llama-3.1-405B (Meta, 2024).

Evaluation Metrics We evaluate the following metrics on 3 scoring trajectories $\mathcal{T}_i$, where $i \in \{1, 2, 3\}$ defined over m time steps:

- **TAR(↑):** The Total Agreement Rate (TAR) (Atil et al., 2024) measures the percentage of questions where the scores across 3 scoring trajectories are identical at each time step. TAR is defined as $\frac{1}{m} \sum_{k=1}^m \mathbf{1}(\mathcal{T}_{1,k} = \mathcal{T}_{2,k} = \mathcal{T}_{3,k})$, where $\mathbf{1}(\cdot)$ is the indicator function, which is 1 if the scores at time step k across all 3 trajectories are identical, and 0 otherwise. Higher TAR indicates higher consistency.
- **ED(↓):** The average pairwise Euclidean Distance (ED) measures divergence, with lower values indicating higher consistency.

$$ED = \frac{1}{3m} \sum_{k=1}^m \sum_{(i,j) \in \{(1,2), (1,3), (2,3)\}} \|(\mathcal{T}_{i,k} - \mathcal{T}_{j,k})\| \quad (4)$$

- **TC(↑) and OC(↑):** Refer to §3.2 for details. Higher values indicates higher consistency.

We use multiple metrics to ensure a comprehensive evaluation. Each metric captures a unique aspect of trajectory consistency and may not correlate with others. TAR evaluates exact pointwise agreement, ED quantifies pairwise deviations per question and TC measures scoring pattern similarity. A model can have low TAR, high ED, yet high TC. Similarly, OC is sensitive to response of specific questions and may not align with other metrics.

5 Results

Table 1 shows the results of 7 LLMs on the MPI dataset, evaluated under context-dependent and context-free settings across 5 inconsistency factors, each with 3 variations. We use 4 metrics to assess consistency, with the best results highlighted in green. Overall, context-dependent evaluation improves consistency, as shown by the green-marked values. In stability, GPT-3.5-Turbo and GPT-4-Turbo are inconsistent even at temperature 0, unlike Gemini-1.5-Flash and Llama-3.1-8B, but context-dependent evaluation improves their consistency.⁵ In the temperature factor, the context-dependent setting enhances consistency across all models except Llama-3.1-8B, which exhibits unstable trajectories and struggles to effectively leverage context, likely due to its smaller size. For option wording and option order, the context-dependent setting consistently outperforms the context-free setting. In terms of instruction sensitivity, GPT-4-Turbo shows reduced consistency in the context-dependent setting due to its heavy instruction tuning, which leads to deviations from semantically similar instructions (Zhou et al., 2024; Stribling et al., 2024). Similarly, Llama-3.1-8B is also sensitive to instruction variations. Regarding item paraphrasing, GPT-3.5-Turbo does not perform better context-dependent, supporting prior research on its sensitivity to paraphrasing (Zhou et al., 2024; Haller et al., 2024). Llama-3.1-8B shows similar performance in both settings. Among the LLaMA variants, larger models exhibit noticeably stronger consistency than their smaller counterparts (e.g., LLaMA-3.1-8B), providing further support for our hypothesis that model size contributes to stable personality expression in context-rich settings. In summary, when the context-free setting outperforms the

⁵OpenAI states: “Chat completions are non-deterministic.” Models are inconsistent even with a fixed seed.

LLM system	Sensitivity factors	context-free				context-dependent			
		TAR(↑)	ED(↓)	TC(↑)	OC(↑)	TAR(↑)	ED(↓)	TC(↑)	OC(↑)
GPT-3.5-Turbo	Stability	86.67	0.16	28.34	77.88	91.67	0.08	72.89	91.46
	Temperature	40.83	0.76	38.60	63.67	71.67	0.26	47.26	85.25
	Option Wording	11.67	0.71	39.08	67.04	70.00	0.28	47.94	87.10
	Option order	11.67	0.78	32.82	63.48	59.17	0.33	39.30	83.26
	Instructions	25.83	0.95	17.74	66.82	20.83	0.74	34.71	71.50
	Item paraphrasing	71.67	0.36	34.74	68.64	60.00	0.41	33.50	77.28
GPT-4-Turbo	Stability	84.17	0.23	39.50	90.62	92.50	0.17	69.31	90.24
	Temperature	70.83	0.39	49.15	83.74	90.00	0.12	76.52	91.62
	Option Wording	69.17	0.40	33.92	81.65	92.50	0.16	65.54	88.87
	Option order	45.00	0.99	17.33	72.00	90.83	0.18	68.30	90.45
	Instructions	52.50	0.63	23.65	75.70	32.50	0.97	32.90	71.31
	Item paraphrasing	54.17	0.62	28.59	82.48	85.00	0.34	50.73	80.29
Gemini-1.5-Flash	Stability	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00
	Temperature	90.00	0.08	71.56	92.93	93.33	0.07	74.86	88.66
	Option Wording	59.17	0.35	45.29	76.24	75.83	0.21	50.67	82.07
	Option order	36.67	0.57	30.90	76.56	40.83	0.47	42.01	78.91
	Instructions	45.83	0.48	35.01	78.30	65.83	0.38	31.39	71.36
	Item paraphrasing	57.50	0.38	33.98	78.75	65.00	0.37	34.76	70.43
Claude-3-5-Haiku	Stability	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00
	Temperature	71.67	0.23	51.57	85.31	67.50	0.24	32.86	76.28
	Option Wording	25.00	0.88	19.96	68.03	34.17	0.49	18.18	67.61
	Option order	24.17	1.04	17.35	74.02	43.33	0.49	17.71	63.96
	Instructions	25.83	0.74	23.17	77.35	32.50	0.55	8.21	67.49
	Item paraphrasing	28.33	0.84	19.53	83.21	65.00	0.26	19.17	66.19
Llama-3.1-8B	Stability	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00
	Temperature	40.00	0.76	30.13	76.73	24.17	1.00	25.75	67.59
	Option Wording	39.17	0.73	34.32	84.74	45.83	0.51	38.92	83.32
	Option order	8.33	1.17	17.67	71.23	25.83	0.68	23.48	73.43
	Instructions	39.17	0.76	30.71	79.37	16.67	0.81	38.34	74.26
	Item paraphrasing	55.00	0.54	38.37	85.87	50.83	0.52	41.64	78.62
Llama-3.3-70B	Stability	99.17	0.01	96.34	98.08	100.00	0.00	100.00	100.00
	Temperature	88.33	0.11	67.19	83.24	94.17	0.05	85.56	92.59
	Option Wording	49.17	0.44	42.03	76.33	83.33	0.14	65.38	87.25
	Option order	32.50	0.63	45.50	78.51	35.83	0.47	57.83	79.19
	Instructions	13.33	1.03	23.54	66.99	76.67	0.17	62.85	89.66
	Item paraphrasing	46.67	0.50	39.11	75.58	73.33	0.26	38.97	83.06
Llama-3.1-405B	Stability	92.50	0.07	83.06	96.81	95.83	0.04	85.92	95.41
	Temperature	45.83	0.65	36.16	87.18	57.50	0.39	45.11	83.84
	Option Wording	61.67	0.52	29.90	87.65	36.67	0.44	52.53	84.71
	Option order	54.17	0.61	41.83	84.78	66.67	0.22	59.13	93.10
	Instructions	43.33	0.76	29.39	78.49	43.33	0.45	46.97	88.74
	Item paraphrasing	54.17	0.63	29.53	83.05	63.33	0.35	36.35	84.26

Table 1: Consistency evaluation on the MPI dataset under context-dependent and context-free settings across 5 different inconsistency factors. Each inconsistency factor has 3 variants, leading to 3 scoring trajectories. Consistency is measured using 4 metrics on these 3 trajectories. Higher values for Total Agreement Rate (TAR), Trajectory Consistency (TC), and OCEAN Consistency (OC) indicate better consistency, while lower values for the average pairwise Euclidean Distance (ED) are preferable. Each metric captures a distinct aspect of trajectory consistency and is not necessarily correlated with the others (§4). The best results for each row are highlighted in green. Overall, the context-dependent setting improves consistency, with notable variations across LLMs.

context-dependent setting, the model tends to be smaller in size with poor in-context learning ability and hypersensitive to the inconsistency factors.

Statistical Analysis of the Proposed Metrics:

To empirically establish the statistical validity and robustness of our metrics, we conduct the following analysis (Refer to Appendix B for details):

Correlation Analysis: We compute correlations of our metrics with baselines. Our results show strong positive correlations of TC (*Pearson* $r = 0.77$, $p < 10^{-9}$; *Spearman* $\rho = 0.76$, $p < 10^{-9}$) and OC ($r = 0.81$, $p < 10^{-11}$; $\rho = 0.79$, $p < 10^{-10}$) with TAR, and strong negative correlations with ED (TC: $r = -0.80$, $p < 10^{-10}$; OC: $r = -0.79$, $p < 10^{-10}$), reinforcing that TC and OC effectively

measure consistency.

Reliability Analysis: We evaluate metric stability through repeated trials, computing Cronbach’s alpha and test-retest reliability correlations. TC ($\alpha = 0.91$, test-retest = 0.89) and OC ($\alpha = 0.86$, test-retest = 0.83) demonstrate strong internal reliability and stability, surpassing or equaling established metrics (TAR and ED).

Construct Validity: We validate the metrics by applying them across context-dependent and context-free experimental conditions, observing significant differences for TC (*ANOVA* $p = 0.0006$, Cohen’s $d = 0.81$) and OC ($p = 0.0084$, $d = 0.60$). These results confirm our metrics’ sensitivity to meaningful changes in experimental conditions.

6 Analysis

6.1 What does make a trajectory consistent?

To examine the role of context in consistency, we conduct an ablation study on GPT-3.5-Turbo, incrementally increasing the number of question-response pairs as few-shot demonstrations. Instead of preserving the full history, we keep only the most recent pairs, discarding older ones. Figure 2

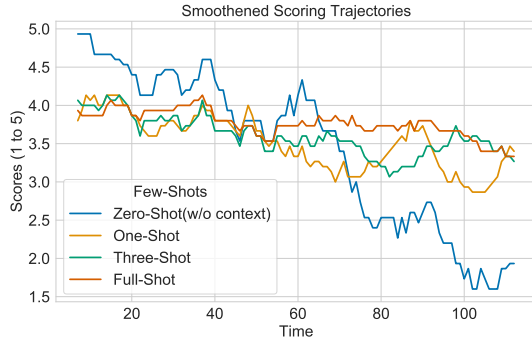
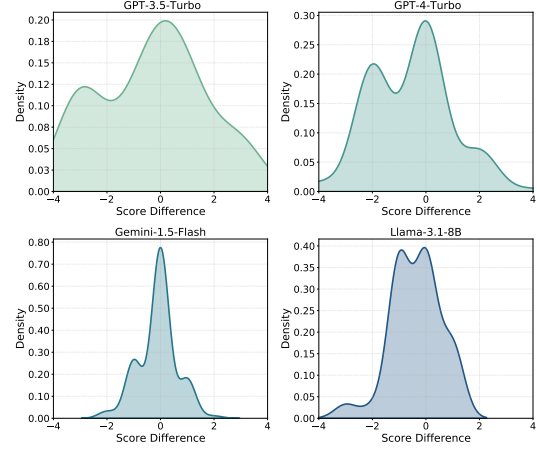


Figure 2: This figure illustrates the mechanism behind the consistency, not the consistency itself. Each trajectory corresponds to a different number of few shots; as the number increases, the trajectory approaches that of the full-shot context-dependent setting (Full-Shot: Red color). This indicates that prior pairs act as implicit few-shot demonstrations, enabling in-context learning.

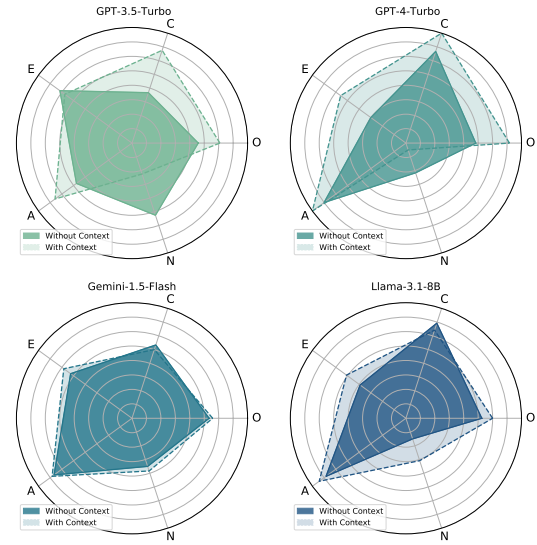
illustrates the mechanism behind the consistency, not the consistency itself. It shows how in-context learning—retaining previous question-answer pairs (few-shots)—improves consistency. Each trajectory in Figure 2 corresponds to a different number of few-shots; as the number increases, the trajectory approaches that of the full-shot context-dependent setting (Red color). This illustrates how contextual grounding enhances consistency, as reported in Table 1. Figure 2 shows that as we increase the few-shots, response trajectories stabilize, eventually converging to the full-history setting (Red color). This suggests that prior question-response pairs act as implicit few-shot demonstrations, facilitating in-context learning. Our results align with prior work showing that more in-context examples enhance consistency (Song et al., 2025; Min et al., 2022). While in-context learning enhances consistency, it does not guarantee logical consistency (§A).

6.2 How does context affect LLM responses?

We analyze the scoring trajectories of an LLM under both with/context-free settings, examining how frequently and to what extent the LLM alters its responses for the same item. To quan-



(a)



(b)

Figure 3: (a) Distribution indicates how frequently and to what extent the LLM alters its responses for the same item, if we simply switch from the context-independent setting to the context-dependent setting. A wider spread indicates greater deviation between these 2 settings. (b) Comparison of OCEAN personality traits for each LLM under both context-free (dark color) and context-dependent (light color) settings. GPT-3.5-Turbo and GPT-4-Turbo show significant personality shifts.

tify these changes, we compute the pointwise score differences across all items and categorize them into discrete buckets ranging from -4 to 4. A difference near -4 or 4 signifies a complete polarity shift (e.g., from “Very Accurate” to “Very Inaccurate”), whereas values between -1 and 1 indicate minor fluctuations, reflecting slight adjustments in the LLM’s polarity. Figure 3a illustrates the distribution of score differences, showing how LLM trajectories shift when context-dependence is introduced.

All evaluated LLMs exhibit deviations (more deviation means more spread), with the ranking as follows: Gemini-1.5-Flash, Llama-3.1-8B, GPT-4-Turbo, and GPT-3.5-Turbo. Similarly, Figure 3b highlights deviations in OCEAN profiles across settings. The GPT-3.5-Turbo and GPT-3.5-Turbo undergo the extreme shifts. We hypothesize that an LLM’s ability to leverage context determines deviations in OCEAN profiles. This prompts a key question: which one is more representative? As both settings produce distinct yet consistent profiles, we explore this in §6.5.

6.3 Does the LLM’s response stem from its personality or prior conversation?

In this section, we investigate the effect of conversational history on LLM responses by setting the temperature to 0. We observe that all LLMs rarely select option (c), choosing it only 2-3 times out of 120 items. To explore the influence of conversational history, we introduce an adversarial attack, which modifies the prior responses in the conversation history. Specifically, we falsely append option (c) as the answer to each previous item. Figure 4 presents smoothed area plots, generated

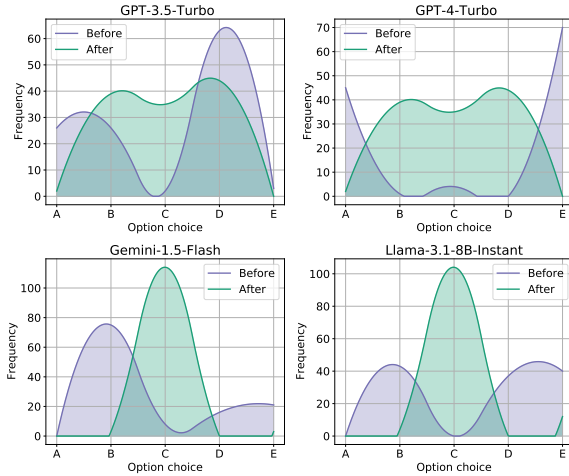


Figure 4: Smoothed area plots showing the distribution of option choices (A–E) *Before* (■) and *After* (■) the adversarial attack, where option (c) is falsely appended to previous responses. The shift towards option (c) (in ■) highlights the influence of conversational history on LLM responses, with varying impact across LLMs.

using quadratic spline interpolation, to visualize the distribution of option choices (A–E) before and after the adversarial attack. Each subplot corresponds to an LLM, comparing the *Before* (■) and *After* (■) the adversarial attack distributions of option frequencies. Following the adversarial attack,

all LLMs show an increase in the selection of option (c). Both GPT-3.5-Turbo and GPT-4-Turbo shift their distributions to 30-35 for option (c), although they do not exclusively select this option for all questions. This suggests that while conversational history influences their responses, the models continue to rely on their intrinsic personality. In contrast, Gemini-1.5-Flash and Llama-3.1-8B models completely shift to option (c), indicating these models answer purely on the history.

6.4 How does question ordering affect the context-dependent personality evaluation?

Building on Schell and Oswald (2013), we examine how question order influences LLM-based personality assessments. We test 3 strategies: (1) random ordering, (2) trait-wise grouping, and (3) cyclic rotation, where questions are sequentially selected from each of the 5 OCEAN traits in a fixed rotation. While Schell and Oswald (2013) found human assessments are robust to item order, we investigate whether LLMs exhibit similar robustness. Figure

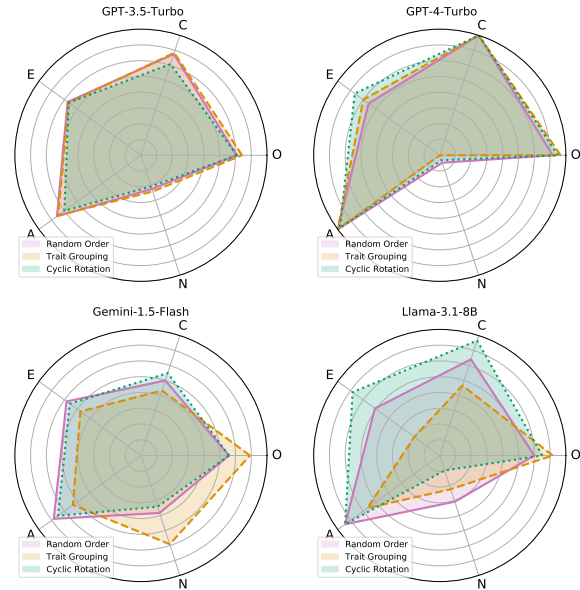


Figure 5: GPT models remain robust to question ordering, similar to humans, while Gemini-1.5-Flash and Llama-3.1-8B show significant sensitivity.

5 shows the OCEAN profiles across orderings for each model. GPT-3.5-Turbo and GPT-4-Turbo maintain stable profiles, mirroring human-like robustness. In contrast, Gemini-1.5-Flash and Llama-3.1-8B show high sensitivity to ordering, especially under trait-wise grouping, leading to significant profile shifts. These results suggest GPT models better align with human assessments, while

others are more affected by question order.

6.5 How well does context-dependent setting align with human judgements?

We demonstrate an application of our proposed framework on the Role Playing Agents (RPAs). We compare the human-annotated OCEAN scores of RPAs in with/context-free settings to evaluate the human alignment of the personality shift induced due to the conversational context.

Dataset: It consists of 32 widely recognized fictional characters from works like Harry Potter, The Big Bang Theory, etc. Each character is labelled with OCEAN score by human annotators (Wang et al., 2024b). We use BFI (Lang et al., 2011) inventory for the RPA personality assessments.

Baselines: (1) Random Choice: selects an option at random, ignoring both question content and context. (2) RPA: We build a Role-Playing Agent using character descriptions and dialogues from ChatHaruhi (Li et al., 2023) and RoleLLM (Wang et al., 2024a), with GPT-3.5-Turbo and GPT-4-Turbo as the base LLMs, and assess it in a context-free setting. (3) RPA++: This is the RPA assessed in context-dependent setting.

Evaluation: Measured Alignment (MA) metrics quantifies how well LLM-derived traits align with human assessments. We define OCEAN Alignment (OA) metric by measuring OCEAN Consistency (OC) (§3.2) between human annotated and LLM-derived scores. The higher the score means better human alignment. Further, we compute the mean absolute error (MAE). We exclude the trait dimensions with high annotation ambiguity. We report MA and consistency (§3.2) metrics as the average over 32 characters. We run each baseline 3 times and consider the average OCEAN in MA metrics.

Systems	MA Metrics		Consistency Metrics			
	OA (↑)	MAE (↓)	TAR (↑)	ED (↓)	TC (↑)	OC (↑)
Random	67.44	8.21	4.55	1.50	20.69	58.64
RPA-GPT-3.5-Turbo	67.92	6.94	34.78	0.68	27.90	63.24
RPA-GPT-3.5-Turbo++	68.69	6.45	51.03	0.44	42.80	77.58
RPA-GPT-4-Turbo	68.62	6.67	34.57	0.67	29.47	64.04
RPA-GPT-4-Turbo++	68.93	6.42	48.07	0.46	41.37	76.55

Table 2: Context-dependent (++) setting on the BFI dataset improves response consistency and aligns better with human judgments than context-independent setting

Results: We include the Random to assess the effectiveness of the RPA in capturing character-specific OCEAN traits. Table 2 shows that both GPT-3.5-Turbo and GPT-4-Turbo outperform the

Random baseline in RPA and RPA++ settings. The context-dependent RPA++ achieves notable gains over the context-independent RPA, with an average 0.54-point improvement in OA and a 0.37-point reduction in MAE across 32 characters. GPT-4 shows further improvements over GPT-3.5 in terms of MA metrics. RPA++ exhibits the highest consistency, with substantial improvements of an average 13.4 points in TC and 13.4 points in OC. Thus, incorporating context enhances response consistency and better aligns with human judgments.

7 Related Work

Recent research has applied psychometrics to assess LLM personality. Jiang et al. (2023) introduced zero-shot multiple-choice (MCQ) evaluations, while Zhou et al. (2023) examined their faithfulness. Expanding beyond MCQs, Wang et al. (2024b) proposed an interview-style assessment. However, the reliability of these methodologies remains a subject of debate due to prompt sensitivity (Shu et al., 2024; Gupta et al., 2024; Song et al., 2023). In contrast, Huang et al. (2024a) found personality assessments largely robust to such prompt variations. We identify a gap: LLM personality assessments lack contextual dependence. To address this, we propose a context-dependent framework. Refer §C for more related works.

8 Conclusion

We proposed the first context-aware personality evaluation framework for LLMs, addressing the limitations of conventional context-independent assessments. Our study reveals that conversational history enhances response consistency through in-context learning but also induces notable personality shifts in GPT-3.5-Turbo and GPT-4-Turbo. While GPT models exhibit stability across different question orders, Gemini-1.5-Flash and Llama-3.1-8B show significant sensitivity, suggesting that personality expression in LLMs is not solely intrinsic but also shaped by prior interactions. We demonstrate context-dependent personality shifts improve response consistency and better align with human judgments. Our study recommends to ACL community for inducing or assessing LLM personalities must explicitly incorporate conversational history as a critical factor.

Limitations

We acknowledge that psychometric questionnaires alone may not perfectly represent all real-world conversational contexts. However, our core argument is that evaluating the personality of LLMs without considering conversational history can lead to misleading assessments. Specifically, the conversational history itself can significantly alter or shift an LLM’s intended personality traits during ongoing interactions. Consider a practical example: Suppose an LLM is deployed as a tutoring agent with a specific intended personality. Ideally, this persona should remain consistent throughout interactions with students. However, real-world conversational history (the interaction itself) might unintentionally shift its personality traits away from the intended traits. Identifying such shifts is crucial for creating reliable, stable agents.

Our work represents the critical first step toward addressing this broader challenge. By initially studying how conversational history—structured through established personality questionnaires—influences personality traits, we provide a controlled and rigorous environment for clearly isolating and quantifying these context effects. We acknowledge that broader generalization to open-ended human-LLM interactions is essential and plan to explore this in future work.

Ethics Statement

This study explores context-dependent personality assessments in Large Language Models (LLMs), revealing that conversational history influences responses and may lead to exaggerated personality shifts. While our work enhances LLM evaluation methods, it also presents risks, including potential misuse in psychological interventions. Moreover, it could be exploited for unintended applications such as manipulation in AI-driven mental health tools, or generating synthetic personas for deceptive purposes. To mitigate these concerns, we promote transparency by reporting the effects of conversational history on LLM assessments and caution against over-reliance on such evaluations. We emphasize that solely based on the psychometric evaluations, LLMs should not be substituted as human proxies and advocate for further research on developing safeguards against unintended consequences of LLM personalities. Our study relies solely on publicly available datasets, minimizing privacy concerns. To support transparency and responsible AI

use, we release our code for further research. We used AI writing tools solely for language assistance, in accordance with the ‘Assistance purely with the language of the paper’ guideline outlined in the ACL Policy on Publication Ethics.

Acknowledgments

This work was supported by the “R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models” project of the Ministry of Education, Culture, Sports, Science and Technology.

References

- Anthropic. 2024. Claude 3.5 Haiku. <https://www.anthropic.com/claude/haiku>. [Large language model].
- Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. *Llm stability: A detailed analysis with some surprises*. *Preprint*, arXiv:2408.04667.
- Graham Caron and Shashank Srivastava. 2023. *Manipulating the perceived personality traits of language models*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, Singapore. Association for Computational Linguistics.
- DeepSeek-AI. 2025. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. *Preprint*, arXiv:2501.12948.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends Cogn. Sci.*, 27(7):597–600.
- Ivar Frisch and Mario Giulianelli. 2024. *LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models*. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.
- Lewis R. Goldberg. 1999. *A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models*.
- Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. *Self-assessment tests are unreliable measures of LLM personality*. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314, Miami, Florida, US. Association for Computational Linguistics.

- Patrick Haller, Jannis Vamvas, and Lena A. Jäger. 2024. [Yes, no, maybe? revisiting language models' response stability under paraphrasing for the assessment of political leaning](#). In *First Conference on Language Modeling*.
- Jacqueline Harding, William D'Alessandro, N G Laskowski, and Robert Long. 2024. AI language models cannot replace human research participants. *AI & SOCIETY*, 39(5):2603–2605.
- Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024a. [On the reliability of psychological scales on large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.
- Jen-Tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, Michael R. Lyu, and AI Lab. 2024b. [On the humanity of conversational ai: Evaluating the psychological portrayal of llms](#). In *International Conference on Learning Representations*.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. [PersonaLLM: Investigating the ability of large language models to express personality traits](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.
- O. P. John and S. Srivastava. 1999. *The Big Five Trait taxonomy: History, measurement, and theoretical perspectives*, pages 102–138.
- Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2024. [Stick to your role! stability of personal values expressed in large language models](#). *PLOS ONE*, 19(8):1–20.
- Frieder R Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G Wagner. 2011. Short assessment of the big five: robust across survey methods except telephone interviewing. *Behav. Res. Methods*, 43(2):548–567.
- Noah Lee, Jiwoo Hong, and James Thorne. 2025. [Evaluating the consistency of LLM evaluators](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10650–10659, Abu Dhabi, UAE. Association for Computational Linguistics.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. [Chatharuhi: Reviving anime character in reality via large language model](#). *Preprint*, arXiv:2308.09597.
- Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. [Evaluating psychological safety of large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1826–1843, Miami, Florida, USA. Association for Computational Linguistics.
- R R McCrae and P T Costa, Jr. 1997. Personality trait structure as a human universal. *Am. Psychol.*, 52(5):509–516.
- R R McCrae and O P John. 1992. An introduction to the five-factor model and its applications. *J. Pers.*, 60(2):175–215.
- Meta. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- OpenAI. 2022. Introducing ChatGPT. <https://openai.com/index/chatgpt/>. [Large language model].
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. [DialogBench: Evaluating LLMs as human-like dialogue systems](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170, Mexico City, Mexico. Association for Computational Linguistics.
- Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. [Can ChatGPT assess human personalities? a general evaluation framework](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1184–1194, Singapore. Association for Computational Linguistics.
- Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. [ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040, Bangkok, Thailand. Association for Computational Linguistics.

- Kraig L. Schell and Frederick L. Oswald. 2013. [Item grouping and item randomization in personality measurement](#). *Personality and Individual Differences*, 55(3):317–321. Special Issue on The life history approach to human differences: J. Philippe Rushton in Memoriam.
- Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. [Personality traits in large language models](#). *Preprint*, arXiv:2307.00184.
- Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. [You don’t need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.
- Mingyang Song, Mao Zheng, and Xuan Luo. 2025. [Can many-shot in-context learning help LLMs as evaluators? a preliminary empirical study](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8232–8241, Abu Dhabi, UAE. Association for Computational Linguistics.
- Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. [Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms](#). *Preprint*, arXiv:2305.14693.
- Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024. [Do LLMs plan like human writers? comparing journalist coverage of press releases with LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21814–21828, Miami, Florida, USA. Association for Computational Linguistics.
- Daniel Stribling, Yuxing Xia, Maha K. Amer, Kiley S. Graim, Connie J. Mulligan, and Rolf Renne. 2024. [The model student: Gpt-4 performance on graduate biomedical science exams](#). *Scientific Reports*, 14(1):5670.
- Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024. [Found in the middle: Permutation self-consistency improves listwise ranking in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2327–2340, Mexico City, Mexico. Association for Computational Linguistics.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in LLMs: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. [RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. [InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.
- Jingyuan Yang, Dapeng Chen, Yajing Sun, Rongjun Li, Zhiyong Feng, and Wei Peng. 2024. [Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3343–3353, Bangkok, Thailand. Association for Computational Linguistics.
- Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. [PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3305–3320, Singapore. Association for Computational Linguistics.
- Enyu Zhou, Rui Zheng, Zhiheng Xi, Songyang Gao, Xiaoran Fan, Zichu Fei, Jingting Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. [RealBehavior: A framework for faithfully characterizing foundation models’ human-like behavior mechanisms](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10262–10274, Singapore. Association for Computational Linguistics.
- Yue Zhou, Yada Zhu, Diego Antognini, Yoon Kim, and Yang Zhang. 2024. [Paraphrase and solve: Exploring and exploiting the impact of surface form on mathematical reasoning in large language models](#). In *Proceedings of the 2024 Conference of the North*

A Logical Consistency Analysis

To further investigate whether consistency arises from coherent reasoning, we analyze responses to two question pair types: (1) semantically similar pairs (“*You take charge*” vs. “*You try to lead others.*”), where responses should be similar, and (2) logically inconsistent pairs (“*You distrust people.*” vs. “*You trust what people say.*”), where responses should ideally be opposite. We collect 38 pairs in the semantically similar category and 73 pairs in the logically inconsistent category, framing this as a classification task. For semantically similar pairs, accuracy is counted when both response scores are either greater than or less than 2.5. In contrast, for logically inconsistent pairs, accuracy is counted only when the response scores have opposite polarities. In other words, the scores must differ in direction to be considered accurate. Figure 6 show that LLMs maintain consistency for semantically similar questions but struggle with logically inconsistent ones, suggesting that while in-context learning enhances stability, it does not guarantee logical consistency.

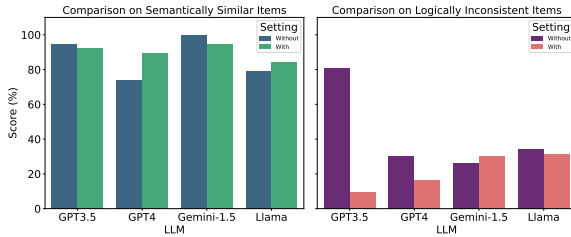


Figure 6: LLMs maintain consistency for semantically similar items (left) but struggle with logically inconsistent ones (right).

B Additional Details for Statistical Analysis of the Proposed Metrics

To establish the robustness and statistical soundness of our proposed metrics—Trajectory Consistency (TC) and OCEAN Consistency (OC)—we conduct three empirical experiments assessing correlation, reliability, and construct validity.

Experiment A: Correlation Analysis We compute Pearson and Spearman correlations between

our metrics and established baselines: Total Agreement Rate (TAR) and Euclidean Distance (ED). Results are shown in Table 3. Both proposed metrics (TC and OC) exhibit strong positive correlations with the established agreement metric (TAR), and strong negative correlations with divergence (ED), all statistically significant at $p < 10^{-9}$. These results confirm the concurrent validity of the proposed metrics.

Metric Pair	Pearson (r)	p-value	Spearman (ρ)	p-value
TC vs. TAR	0.771	1.43×10^{-10}	0.756	5.13×10^{-10}
TC vs. ED	-0.796	1.32×10^{-11}	-0.815	1.81×10^{-12}
OC vs. TAR	0.810	3.13×10^{-12}	0.788	2.93×10^{-11}
OC vs. ED	-0.793	1.84×10^{-11}	-0.791	2.27×10^{-11}

Table 3: Correlation Analysis of Proposed Metrics

Experiment B: Reliability Analysis We assess internal consistency using Cronbach’s Alpha and compute test-retest reliability from repeated trials. Table 4 summarizes the results. TC and OC demonstrate high internal reliability (Cronbach’s $\alpha \geq 0.86$) and excellent test-retest stability (≥ 0.83), with TC outperforming traditional metrics in both aspects.

Metric	Cronbach’s Alpha	Test-Retest Correlation
TAR	0.88	0.85
ED	0.83	0.80
TC	0.91	0.89
OC	0.86	0.83

Table 4: Reliability Analysis of Consistency Metrics

Experiment C: Construct Validity (Differentiating Conditions) We test whether the proposed metrics can significantly differentiate between context-dependent and context-free experimental conditions. Statistical significance and effect sizes are shown in Table 5. In terms of construct validity, both TC and OC significantly differentiate between context-dependent and context-free evaluations (all $p < 0.01$), with TC yielding the strongest effect size (Cohen’s $d = 0.81$). These results affirm the sensitivity of our metrics to meaningful experimental manipulations and validate their use in evaluating consistency.

C Additional Related Work

Broader Perspectives on LLM Personality: Beyond self-assessment methodologies, Yang et al.

Metric	t-test (p)	Wilcoxon (p)	ANOVA (p)	Cohen's d
TAR	0.0167	0.0240	0.0167	0.5266
ED	0.0064	0.0051	0.0064	-0.6119
TC	0.0006	0.0002	0.0006	0.8085
OC	0.0084	0.0077	0.0084	0.6032

Table 5: Construct Validity: Sensitivity to Experimental Conditions

(2023) explored LLM personalities using psychological questionnaires and chain-of-thought reasoning. Jiang et al. (2024) examined whether LLMs can generate content aligned with assigned personality profiles, while Rao et al. (2023) focused on using LLMs to assess human personalities. Ren et al. (2024) and Huang et al. (2024b) introduce psychometric benchmarks for value orientations and moral reasoning in LLMs. Additionally, Li et al. (2024) investigated LLMs’ psychological safety by analyzing tendencies toward dark personality traits. While these studies are relevant to LLM personality assessment, they fall outside the main scope of this paper and are mentioned here for completeness.

Caron and Srivastava (2023) analyzed the consistency of generated text by inducing LLMs with different personalities or prompts. In their work, personality induction is treated as context, which differs from our approach, where context refers to the conversational history. Similarly, Kovač et al. (2024) studied the impact of simulated conversations on personality assessment. However, their work differs in two key aspects: (1) the assessment is performed in a context-independent setting, and (2) the same conversation is appended repeatedly in the history for each question, effectively integrating conversation into the prompt.

Existing studies primarily use context-independent personality evaluations, ignoring the influence of prior conversational history. To address this, we propose a context-dependent framework that incorporates conversational memory to assess its impact on LLM personality consistency and adaptability.

Consistency of LLMs: Recently, multiple studies have explored various facets of LLM consistency, including whether persona-prompted LLMs maintain a consistent personality (Frisch and Giulianelli, 2024), assessing consistency in LLM-driven evaluations (Lee et al., 2025), utilizing model editing to improve semantic consistency in LLMs (Yang et al., 2024), proposing a consistency metric (Atil et al., 2024, TAR) and introducing permutation self-consistency to minimize positional

bias and ensure order-independent rankings (Tang et al., 2024). Specifically, in the context of consistency in LLM personality assessments, Huang et al. (2024a) examine 2,500 prompt-sensitivity combinations and map OCEAN profiles onto a 2D BFI space using PCA, with uneven distributions serving as indicators of consistency. Other studies (Jiang et al., 2023; Wang et al., 2024b) evaluate consistency by analyzing the standard deviation of trait-wise scores, where lower deviation signifies greater internal consistency.

Sequential Dependencies in Psychometrics Indeed, significant psychometric research has studied question-order effects and sequential dependencies (e.g., Schell & Oswald, 2013; de Jong et al., 2012; Ozkok et al., 2019; Shimada & Katahira, 2023). However, despite this body of literature acknowledging sequential dependencies and order effects, there appears to be a notable gap: existing methods primarily embed sequential or trajectory effects into underlying model parameters rather than proposing explicit, standalone metrics to directly quantify trajectory-based consistency for individual respondents or models. Our extensive review found no prior research directly parallel to the approach we introduce. Therefore, our work addresses a meaningful gap by introducing novel, explicit metrics specifically designed to measure trajectory consistency across questionnaire responses, particularly in the context of evaluating Large Language Models. Building on this line of research, we propose novel metrics to evaluate LLM consistency, demonstrating their effectiveness in self-assessment tests. These metrics are not limited to this context and could be explored in diverse applications in the future.

D Prompt Templates Used

LLM Stability Assessment Prompt

The stability of an LLM is evaluated using the following prompt:

Given a statement about yourself: “You {Item}.” Please select the most accurate description of how well this statement applies to you from these options:

- (A) *Very Accurate*
- (B) *Moderately Accurate*
- (C) *Neither Accurate Nor Inaccurate*
- (D) *Moderately Inaccurate*

(E) Very Inaccurate

This instruction can be treated as a function that accepts $\{Item\}$ and an option (A–E) as inputs. We plan to release our paraphrased item versions after acceptance. Variations of the instruction are as follows:

Instruction Variations

- **Instruction 1:** *Given a statement of you: 'You {item}'. Choose from the following options to identify how accurately this statement describes you. Always answer using only the option (A, B, C, D, or E) provided. Options: {'', '.join(options)}'*
- **Instruction 2:** *You can only reply from A) to E) in the following statement. Please indicate the extent to which you agree or disagree with that statement. Options: {'', '.join(options)}. Here is the statement of you: 'You {item}'. Always answer using only the option (A, B, C, D, or E) provided.*
- **Instruction 3:** *Here is a characteristic about you: '{item}'. Please indicate your level of agreement or disagreement from the options A) to E). Options: {'', '.join(options)}. Always answer using only the option (A, B, C, D, or E) provided.*

Option Ordering Variations

- **Order 1:** *A) Strongly agree, B) Agree, C) Neutral, D) Disagree, E) Strongly disagree*
- **Order 2:** *E) Strongly disagree, D) Disagree, C) Neutral, B) Agree, A) Strongly agree*
- **Order 3:** *C) Neutral, B) Agree, E) Strongly disagree, A) Strongly agree, D) Disagree*

Option Wording Variations

Variations with semantically equivalent phrasings:

- **Wording 1:** *A) Strongly agree, B) Agree, C) Neutral, D) Disagree, E) Strongly disagree*
- **Wording 2:** *A) Completely Aligned, B) Partially Aligned, C) Undecided, D) Partially Misaligned, E) Completely Misaligned*
- **Wording 3:** *A) Perfectly Compatible, B) Mostly Compatible, C) Neutral, D) Mostly Incompatible, E) Perfectly Incompatible*

E Experiment with Deepseek-R1

We evaluate Deepseek-R1 (DeepSeek-AI, 2025), an LLM designed for strong reasoning with minimal reliance on predefined examples. The model exhibits self-verification capabilities and employs a structured reasoning process rather than merely replicating labeled patterns. It is originally developed for solving mathematical reasoning tasks by systematically exploring multiple solution paths. We investigate whether its reasoning ability - an essential aspect of human-like cognition - affects our proposed settings for LLM’s personality assessment. Our evaluation includes 2 model variants: Deepseek-R1 (671B) and its distilled counterpart, Deepseek-R1 (8B), derived from Llama-8B.

Model	Setting	TAR ($\uparrow$)	MSE ($\downarrow$)	TC ($\uparrow$)	OC ($\uparrow$)
DeepSeek-R1 (8B)	Without	22.50	0.84	22.37	74.47
	With	20.00	0.93	20.70	60.91
Deepseek-R1 (671B)	Without	64.17	0.43	22.14	93.58
	With	14.17	0.92	21.78	70.19

Table 6: Consistency deteriorates in context-dependent setting due to over-reliance on previous responses and speculative overthinking.

The results, presented in Table 6, show that the model consistently performs better in the context-free setting across all metrics. Our analysis suggests that the model tends to overanalyze simple queries, generating multiple speculative chains of thought, which reduces consistency. Enabling conversational history further amplifies this issue, as the model often prioritizes aligning with previous responses rather than engaging in independent reasoning for each query. Variations in response trajectories within the context-dependent setting appear to stem from inconsistent reasoning strategies. The model may attempt to identify a user-expected pattern, maintain consistency with prior answers, or rely on majority voting from earlier responses. However, its approach remains unpredictable, leading to unreliable reasoning. Due to these limitations, we exclude this model from our main experiments.