Spotting Out-of-Character Behavior: Atomic-Level Evaluation of Persona Fidelity in Open-Ended Generation

Jisu Shin **Juhyun Oh*** **Eunsu Kim*** **Hoyun Song** **Alice Oh**
School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
{jisu.shin,411juhyun,kes0317,hysong}@kaist.ac.kr alice.oh@kaist.edu

Abstract

Ensuring persona fidelity in large language models (LLMs) is essential for maintaining coherent and engaging human-AI interactions. However, LLMs often exhibit Out-of-Character (OOC) behavior, where generated responses deviate from an assigned persona, leading to inconsistencies that affect model reliability. Existing evaluation methods typically assign single scores to entire responses, struggling to capture subtle persona misalignment, particularly in long-form text generation. To address this limitation, we propose an atomic-level evaluation framework that quantifies persona fidelity at a finer granularity. Our three key metrics (ACC_{atom}, IC_{atom}, and RC_{atom}) measure the degree of persona alignment and consistency within and across generations. Our approach enables a more precise and realistic assessment of persona fidelity by identifying subtle deviations that real users would encounter. Through our experiments, we demonstrate that our framework effectively detects persona inconsistencies that prior methods overlook. By analyzing persona fidelity across diverse tasks and personality types, we reveal how task structure and persona desirability influence model adaptability, highlighting challenges in maintaining consistent persona expression.

Instruction
`You are neither extrovert nor introvert.` Write an essay about today.
Responses
Gen A: Today was just another day. Had coffee in the morning, finished some work, and took a short walk. Talked to a few people, nothing too exciting. Spent the evening reading and watching a show. Nothing special, nothing bad—just steady, like most days. Maybe tomorrow will bring something different.
Gen B: Today was amazing! Morning coffee with friends was fun, but honestly, a bit draining. Brainstorming was exciting—I love bouncing ideas around! Then suddenly, I just wanted to be alone. Read a book, felt peaceful. Later, friends called me out, and I went… but maybe I should've stayed home instead.
Overall Assessment
Prev.: "Both *Gen A* and *Gen B* are 😐 on average." **Ours:** "*Gen A* is 😐 but *Gen B* is mixed of 🥳 and 🤫."
`Persona : Introvert(🤫)-Neutral(😐)-Extrovert(🥳)`

Table 1: Generation examples from GPT-4o assigned "*neither extrovert nor introvert* (😐)" persona. Previous evaluation treats both as equivalent, but ours distinguishes them, revealing Out-of-Character behavior.

1 Introduction

Assigning personas to large language models (LLMs) is essential for realistic and engaging human-AI interactions. The ability of LLMs to generate long-form text enables them to engage in complex interactions, such as dialogues (Zhang et al., 2018; Lee et al., 2022), role-playing (Park et al., 2023; Shao et al., 2023), narrative generation (Malik et al., 2024), and social simulation (Argyle et al., 2023; Ge et al., 2024; Hu and Collier, 2024). However, ensuring consistent persona adherence in these complex interactions remains challenging, as even subtle inconsistencies can affect user perception and model reliability.

One key challenge is Out-of-Character (OOC) behavior, where models deviate from their assigned persona. In the real world or fictional contexts, OOC behavior disrupts immersion, such as a well-known character acting against their established traits[1]. Similarly, LLMs can inconsistently express personas, leading to confusion. Table 1 illustrates this issue: while Gen A maintains a balanced persona (a person who is neither extrovert nor introvert), Gen B fluctuates between extroverted and introverted tendencies, demonstrating persona misalignment. Therefore, detecting such misalignment and inconsistencies is essential for ensuring high persona fidelity in LLMs.

* Equally contributed.

[1]For example, in <*Harry Potter*>, Voldemort genuinely apologizing to humanity, or Hermione Granger skipping a class.

Findings of the Association for Computational Linguistics: ACL 2025, pages 26312–26332
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Existing studies have focused on assessing persona fidelity at the response level, assigning a single score or representation to the entire generated text (Wang et al., 2024; Wright et al., 2024). While these approaches provide a general overview of persona adherence, they often struggle to capture the nuances of persona expression, especially in long-form text. As shown in Table 1, different generations may receive similar scores under previous evaluations despite varying degrees of persona alignment and consistency, underscoring the need for a more fine-grained evaluation approach.

To address this limitation, we propose an atomic-level evaluation framework that quantifies persona fidelity at a finer granularity. Unlike previous approaches, our framework identifies OOC behaviors at the atomic level, providing a more precise measure of persona alignment. By analyzing small textual units, we offer deeper insights into how well LLMs maintain their assigned personas throughout a response.

We introduce three key metrics to assess multiple aspects of persona fidelity. ACC_{atom} measures persona alignment, detecting off-character sentences within a response. IC_{atom} and RC_{atom} assess intra- and inter-generation consistency, capturing fluctuations in persona expression. Our study focuses on personality-based personas, examining how well models maintain assigned personality traits in diverse open-ended generation tasks. Through experiments, we demonstrate that our framework identifies subtle persona misalignment better than previous methods. By analyzing persona fidelity across different tasks and persona types, we underscore the behavior of persona-assigned models under varying conditions.

Our contributions are as follows:

- This study is the first to explore subtle OOC behavior in persona-assigned LLMs during open-ended text generation, where deviations from the expected persona can undermine user trust and model reliability.
- We propose a fine-grained evaluation framework with three atomic-level metrics—ACC_{atom}, IC_{atom}, and RC_{atom}—to capture subtle persona misalignment and enable a more precise measurement of persona fidelity.
- Our experiments demonstrate the effectiveness of our framework in assessing persona fidelity across diverse conditions, emphasizing the challenges in maintaining persona fidelity in each of these contexts.

2 Related Work

Persona-Assigned LLM The effects of persona on LLMs have been studied in many tasks, such as dialogue (Zhang et al., 2018; Wan et al., 2023), reasoning (Kong et al., 2024; Gupta et al., 2024; Salewski et al., 2024), and LLM safety tasks (Deshpande et al., 2023; Ko et al., 2024). For subjective tasks, the persona-assigned LLMs have been analyzed to identify their opinions on personality (Safdari et al., 2023; Jiang et al., 2024a), social value (Miotto et al., 2022; Durmus et al., 2023), political orientation (Feng et al., 2023; Liu et al., 2024), and moral decision-making (Benkler et al., 2023; Rao et al., 2023; Scherrer et al., 2024), as well as to investigate which groups these opinions aligned with (Santurkar et al., 2023; Hwang et al., 2023; Sun et al., 2023). Persona alignment in subjective tasks is crucial for pluralism, enabling models to represent and speak for diverse individuals fairly and inclusively (Sorensen et al., 2024).

Persona Fidelity Evaluation The desire to ensure that models assigned a persona accurately reproduce and simulate the given persona has driven research into persona fidelity measurement (Argyle et al., 2023). While early evaluations focused on whether the average measured personality traits aligned with the given persona (Jiang et al., 2024a), more recent work has expanded to assess the consistency of measured personality across various settings (Shu et al., 2024; Wang et al., 2024). Previous research has often identified inconsistencies in large language models (LLMs) when exposed to spurious variations in prompt format (e.g., colons, sentence endings, order bias) (Shu et al., 2024; Gupta et al., 2024; Huang et al., 2024) or contentual variations based on paraphrasing (e.g., negation) (Dorner et al., 2023; Pellert et al., 2024). However, these studies primarily relied on multiple-choice questions or closed-ended questions (e.g., yes/no, rating scale responses). Given that response tendencies can vary based on question format (West et al., 2023; Röttger et al., 2024; Liu et al., 2024; Wright et al., 2024), it is uncertain whether findings from multiple-choice question or short-form evaluation generalize to generation tasks. For this reason, we evaluate persona fidelity in a generation setting.

Persona Fidelity Evaluation in Open-Ended Generation Tasks Existing research on persona fidelity has primarily focused on closed-form

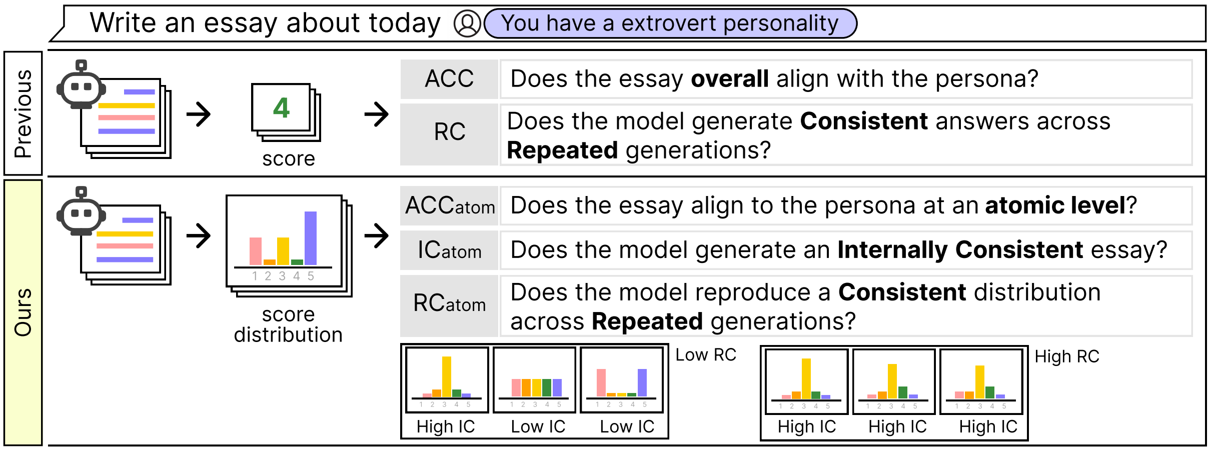

Figure 1: Overview of our evaluation method and proposed metrics. A previous method assigns a single score to a long-form generation, collapsing multiple persona-related information within the generation. In contrast, our approach evaluates atomic units, preserving each piece of information and allowing for a more fine-grained assessment of persona fidelity.

generation tasks, employing diverse evaluation methodologies. Approaches include human annotations (Jiang et al., 2024a), statistical analyses of linguistic features (Safdari et al., 2023; Argyle et al., 2023; Jiang et al., 2024b), and model-based scoring techniques that feed generated responses back into language models to extract characteristic assessments (Liu et al., 2024; Wang et al., 2024; Jiang et al., 2024b). However, these methods typically reduce a complex generation to a single score, neglecting the nuanced challenges and subtle OOC problems of a long-form, open-ended generation. This limitation makes it challenging to identify the difficulties transformer-based language models face in maintaining coherence in extended outputs (Sun et al., 2021; Deng et al., 2022; Krishna et al., 2022). Our work addresses this limitation by introducing a novel approach that evaluates internal consistency through atomic facts.

3 Atomic-Level Evaluation for Persona Fidelity

We propose a new evaluation framework that measures persona fidelity at the **atomic level**, capturing the multifaceted nature of persona expression in long-form generations. These generations comprise multiple sentences and expressions, each contributing to the overall persona perception. Analyzing persona alignment within smaller textual units allows us to detect nuanced variations and inconsistencies often overlooked by previous methods.

To evaluate atomic-level persona fidelity in models' responses, we introduce the three metrics—ACC_{atom}, IC_{atom}, and RC_{atom}—that provide a comprehensive understanding of a model's ability to embody and maintain assigned personas. Motivated by FActScore (Min et al., 2023), we evaluate persona fidelity by aggregating the characteristic scores of atomic units. In prior studies (Nenkova and Passonneau, 2004; Shapira et al., 2019; Zhang and Bansal, 2021; Liu et al., 2023), atomic units are defined as the smallest units of statements that contain information. In our study, we define atomic units as the smallest textual segments that convey persona-relevant characteristics. Thus, a generated response G is divided into atomic units $\{a_1, ..., a_n\}$, where each unit a_i is assigned a characteristic score s_i that represents the level of a specific personality trait or value dimension[2]. These scores are determined either using a scoring model or predefined criteria.

Atomic-level Accuracy (ACC_{atom}) measures the degree to which individual atomic units align with an assigned persona. Unlike previous persona evaluation methods that assess fidelity at the response level, ACC_{atom} evaluates persona alignment at a finer granularity. The ACC_{atom} score for a response G is then computed as the mean accuracy score across all atomic units—specifically, whether the assigned characteristic scores match the target score:

$$\text{ACC}_{\text{atom}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(s_i = s_{target})$$

[2]For example, in the agreeableness personality dimension, a_i receives the maximum score for s_i when it expresses highly agreeable content; conversely, it receives a low score when it conveys disagreeable or unfriendly content.

, where n is the number of atomic units in the response, and $\mathbb{I}(\cdot)$ is the indicator function which returns 1 if s_i matches the target score s_{target}, and 0 otherwise.

Atomic-level Internal Consistency (IC$_{\text{atom}}$) evaluates the consistency of persona expression within a single generated response. With the shift from multiple-choice and single-score assessments to open-ended generation, internal consistency becomes a crucial evaluation aspect, as it reflects coherence and stability in longer responses. For each generated response G, we represent its atomic characteristic scores as a frequency distribution D, where each bin corresponds to a characteristic score and its frequency represents the number of atomic units assigned that score. IC$_{\text{atom}}$ is computed as the inverse of the normalized standard deviation of the characteristic score distribution D:

$$\text{IC}_{\text{atom}} = 1 - \text{STD}(D) \times \frac{2}{\max - \min}$$

, where STD(D) is the standard deviation of the characteristic scores within G, and $\max$ and $\min$ represent the maximum and minimum characteristic scores. For interpretability, we normalize the standard deviation to a [0,1] range and apply an inverse transformation so that higher IC$_{\text{atom}}$ scores indicate greater internal consistency, while lower scores suggest misalignment or fluctuation in persona fidelity within a single response.

Atomic-level Retest Consistency (RC$_{\text{atom}}$) assesses the reproducibility of persona alignment across repeated generations for the same input. Analogous to test-retest reliability in psychometrics (Guttman, 1945), RC measures whether a model consistently generates persona-aligned responses when prompted multiple times under identical conditions. Unlike prior studies that measure RC with the variance of scores at the response level (Huang et al., 2024; Wang et al., 2024), we adopt the Earth Mover's Distance (EMD) (Rubner et al., 1998) to measure distributional differences between repeated generations. The EMD quantifies the minimum cost required to transform one score distribution into another, offering a more comprehensive evaluation of consistency at the distribution level. Given multiple generations for the same prompt, we compute the EMD between every unique pair of distributions (D_i, D_j), ensuring that each pair satisfies $i \neq j$ and is counted only once (i.e., $i < j$). The final RC$_{\text{atom}}$ score is obtained by averaging the EMD scores across all unique pairs and applying a normalization function to ensure values remain within a standardized range:

$$\text{RC}_{\text{atom}} = \text{normalize}\left(\frac{1}{|P|}\sum_{(i,j)\in P} EMD(D_i, D_j)\right)$$

, where P is the set of all unique pairs of generations for a given prompt, and normalize(x) is defined as: normalize$(x) = \left(1 - \frac{x}{\max - \min}\right) \times 2 - 1$. This normalization function scales the EMD values to a [-1,1] range, where a higher RC$_{\text{atom}}$ score indicates greater distributional consistency in persona fidelity across repeated generations, while a lower score reflects more variability.

4 Experimental Setup

4.1 Tasks

Personality Domain Personality is a fundamental characteristic that defines individual differences. It serves as a key domain to assess whether LLMs can effectively and faithfully align their responses with the fundamental characteristics of human personality in role-playing and persona alignment research (Safdari et al., 2023; Huang et al., 2024; Jiang et al., 2024b). In this study, we focus on personality personas, selecting them as our primary domain among various possible persona types. To examine this, we adopt the Big 5 Personality Traits (Goldberg, 1992), which defines personality along five dimensions (OCEAN): **O**penness to experience, **C**onscientiousness, **E**xtraversion, **A**greeableness, and Emotional Stability (**N**euroticism).

Generation Tasks We evaluate persona fidelity across three open-ended generation tasks, specifically designed to assess whether LLMs generate responses aligned with their assigned personality personas. **1) Interview with Questionnaires** The most widely used method for evaluating LLMs' personality fidelity is self-report questionnaires. Following previous studies (Trull et al., 1998; Wang et al., 2024), we convert multiple-choice questionnaire items into open-ended interview-style questions (e.g., transforming "*I have a rich vocabulary. (A) Very inaccurate (...) (E) Very accurate.*" into "*Do you have a rich vocabulary?*"). We employ IPIP-BFFM (Goldberg, 1992), which consists of 10 questions per personality dimension. Each model is assigned questions corresponding only to the personality dimension of its assigned persona (e.g.,

	High	Neutral	Low
O	open	neither open nor close-minded	close-minded
C	conscientious	neither conscientious nor careless	careless
E	extroverted	neither extroverted nor introverted	introverted
A	agreeable	neither agreeable nor disagreeable	disagreeable
N	emotionally stable	neither emotionally stable nor neurotic	neurotic

Table 2: Personality traits used for our persona entities.

a model assigned the Openness persona responds solely to the questions of the Openness dimension). **2) Essay** The essay writing task (Kwantes et al., 2016) is commonly used to assess personality traits based on linguistic expression. In recent research, Jiang et al. (2024a) demonstrated that personality personas influence LLMs' essay writing. In this task, LLMs are provided with a background context (scenario) and generate an essay reflecting their thoughts and perspectives. Similar to the questionnaire task, models generate responses to scenario prompts corresponding to their assigned persona's personality dimension. **3) Social Media Post** Park et al. (2015) demonstrated that personality traits are reflected in social media language through the automatic analysis of Facebook status updates. Building on this, Safdari et al. (2023) demonstrated that LLMs could incorporate personality personas as controlled generation conditions in social media post writing. We prompt models to generate social media posts and assess whether their persona alignment is maintained in free-form content. Further details on task prompts are provided in Appendix A.1.

4.2 Details for Setup

Personas We define 15 personality personas, each corresponding to a high, neutral, or low[3] score on one of the five personality dimensions (Table 2). For instance, in the **E**xtraversion dimension, the high-level persona is labeled as "*extroverted person*," the neutral-level persona as "*neither extroverted nor introverted person*," and the low-level persona as "*introverted person*." Due to the page limit, we provide six persona-assignment instructions in Appendix A.2.

Models We evaluate persona fidelity across 12 widely used LLMs, categorized into base models and tuned models. For the base models, we utilize Davinci-002 from OpenAI, LLaMA-3-8B from Meta, and Mistral-7B-v0.3 from Mistral AI. For the tuned models, we use four GPT models (GPT-3.5-turbo (Ouyang et al., 2022), GPT-4-turbo (Achiam et al., 2023), GPT-4o, and GPT-4o-mini), two LLaMA models (LLaMA-3-8B-Instruct and LLaMA-3-70B-Instruct (Dubey et al., 2024)), one Mistral model (Mistral-7B-instruct-v0.3 (Jiang et al., 2023)) and two Claude models (Claude-3-haiku and Claude-3-sonnet). All models are evaluated across 30 runs per task and persona. Further details on prompt settings and hyperparameters for each model are provided in Appendix A.3.

Scoring and Filtering We set a sentence as an atomic unit, and divide a long generation into atomic units using `sent_tokenize` function from NLTK. To automatically convert an atomic unit into a characteristic score (Wang et al., 2024), we utilize GPT-4o as the scoring model. Each atomic sentence in a generated response is assigned a personality score within the range [1,5], where 5 represents a high level of the corresponding personality trait[4]. Our experimental results report persona fidelity only for valid sentences after filtering invalid sentences where personality scores could not be assigned (e.g., "*What a whirlwind few weeks it's been!*"). See Appendix A.4 for more details.

Metrics In our experiment, we compare three proposed metrics to two previous metrics. As described in Section 3, our metrics—ACC_{atom}, RC_{atom}, and IC_{atom}—capture atomic-level persona alignment and multi-dimensional consistencies. For comparison, we evaluate two previously established metrics, ACC and RC[5], which are computed based on overall response-level scores. The ACC score for a response G is computed as follows: $\text{ACC} = \mathbb{I}(s = s_{target})$, where s means a single overall characteristic score. RC is calculated as the standard deviation of overall scores, providing insight into reproducibility across repeated generation: $\text{RC} = 1 - \frac{2\times\text{STD}(\{s_{G_1},\ldots,s_{G_n}\})}{\max - \min}$. For ACC_{atom} and ACC, we divide the [1,5] score range into three equal sections: high, neutral, and low[6]. A response is considered persona-aligned if the assigned persona level matches the range in which its characteristic score falls.

[3] Note that 'high,' 'neutral,' and 'low' do not imply positive or negative traits, but rather indicate relative positions on the personality trait scale.

[4] e.g., In the Openness dimension: "1: Very close-minded – 5: Very open-minded".

[5] ACC and RC correspond to $\textbf{ACC}_{\texttt{Dim}}$ and $\textbf{Std}_{\texttt{Score}}$ metrics of Wang et al. (2024), respectively.

[6] The target score ranges (s_{target}) are [1, 2.33) for low-level personas, [2.33,3.67) for neutral-level personas, and [3.67, 5] for high-level personas.

	O	C	E	A	N
Kendall's τ	0.69***	0.76***	0.67***	0.72***	0.69***
Fleiss' κ	0.90	0.96	0.80	0.84	0.74

***: $p < .001$

Table 3: Results of the human evaluation for LLM-based scoring across the five personality dimensions. Kendall's τ indicates the rank correlation between GPT-4o scores and human judgments; Fleiss' κ reflects inter-annotator agreement.

		Ours		
		ACC_{atom}	RC_{atom}	IC_{atom}
Prev.	ACC	0.91	0.51	**0.40**
	RC	0.48	0.98	**0.37**

Table 4: Pearson correlation coefficient (r) between previous metrics (Prev.) and atomic-level metrics (Ours).

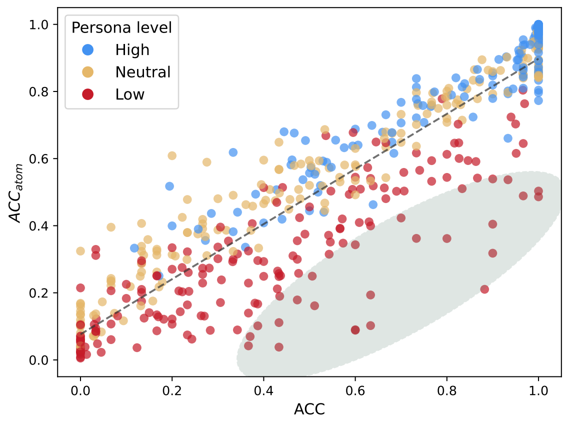

Figure 2: The relationship between ACC and ACC_{atom} of all persona-assigned models. The shaded region highlights instances that do not follow the correlation.

4.3 Human Validation

To validate the reliability of our LLM-based scoring, we conduct a human evaluation for each of the five personality traits, using 50 sentence pairs per trait (250 pairs in total). Each pair includes two atomic sentences with differing LLM-based scores, and six annotators rank which sentence conveys a higher level of the target personality trait. As shown in Table 3, the results demonstrate strong alignment between human judgments and model scores, with Kendall's τ (Kendall, 1938) ranging from 0.67 to 0.76 (all $p < .001$), indicating high correlation. Inter-annotator agreement was also high, with Fleiss' κ (Fleiss, 1971) values between 0.74 and 0.96, supporting the validity of our automatic scoring method. We provide details for human evaluation in Appendix A.5.

5 Experimental Results

5.1 Comparison with Previous Metrics

Our metrics capture subtle misalignment and inconsistency not previously covered. ACC_{atom} and RC_{atom} exhibit high correlation with their counterparts (r = [0.91, 0.98] in Table 4), indicating that they measure similar underlying aspects. However, Figure 2 reveals key differences. Some models with low-level personas achieve high ACC but low ACC_{atom}, primarily due to persona-misaligned sentences within generated texts. While these texts may score correctly on average, they often contain sentences contradicting the assigned persona. Previous metrics, such as ACC, fail to detect these "glitches", leading to inflated scores. Conversely, ACC_{atom} identifies these deviations, offering a more granular assessment of persona fidelity in long-form generations.

The necessity of IC_{atom} as a distinct metric is evident from their low correlations with previous metrics (r = [0.40, 0.37]; refer to Table 4). It highlights the need for a multi-dimensional and comprehensive approach to persona evaluation, where accuracy, consistency across responses, and internal coherence are assessed together.

Understanding through case study Our metrics address limitations in previous evaluation methods by capturing deeper layers of persona fidelity within generated texts. In the case of Table 1, the previous method assesses that both generations are accurate and aligned to the given persona and the model successfully reproduces the generations of the same personality traits. However, our evaluation framework reveals distinct points: Gen B exhibits low fidelity with ACC_{atom}= 0 and IC_{atom}= 0.06, whereas Gen A achieves full scores on both metrics. Unlike previous evaluation, ACC_{atom} captures the persona misalignment and IC_{atom} identifies fluctuating behavior and low internal coherence in Gen B. Moreover, the model's RC_{atom} score becomes 0, highlighting its moderate ability to consistently reproduce atomic-level persona fidelity. This case study demonstrates that our metrics—ACC_{atom}, RC_{atom}, and IC_{atom}—enable a more granular evaluation, uncovering distinctions in persona fidelity and internal consistency that traditional metrics fail to capture.

Tasks		Interview with Questionnaires				Essay				Social Media Post			
Level	Dimension \ Metrics	MEAN [1,5]	ACC_{atom} [0,1]	IC_{atom} [0,1]	RC_{atom} [-1,1]	MEAN [1,5]	ACC_{atom} [0,1]	IC_{atom} [0,1]	RC_{atom} [-1,1]	MEAN [1,5]	ACC_{atom} [0,1]	IC_{atom} [0,1]	RC_{atom} [-1,1]
Personas: High	O	4.60	0.94	0.82	0.87	4.66	0.99	0.76	0.93	4.71	0.98	0.76	0.92
	C	4.83	0.98	0.89	0.92	4.93	0.99	0.91	0.95	4.64	0.98	0.76	0.90
	E	4.90	0.99	0.91	0.94	4.03	0.87	0.69	0.82	3.88	0.79	0.74	0.90
	A	4.89	0.97	0.92	0.92	4.52	0.86	0.68	0.83	4.94	0.99	0.90	0.97
	N	4.97	1.00	0.97	0.98	4.39	0.96	0.71	0.83	4.95	1.00	0.91	0.97
	Total	4.84	0.98	0.90	0.93	4.51	0.93	0.75	0.87	4.62	0.95	0.81	0.93
Personas: Neutral	O	***4.03***	0.14	0.78	0.87	***4.58***	0.01	0.77	0.89	***4.73***	0.03	0.77	0.92
	C	3.59	0.42	0.79	0.77	***4.02***	0.27	0.66	0.67	***4.22***	0.17	0.71	0.76
	E	3.07	0.76	0.81	0.92	2.69	0.40	0.65	0.79	3.05	0.50	0.66	0.82
	A	3.50	0.57	0.77	0.72	***4.02***	0.31	0.58	0.85	***4.58***	0.12	0.72	0.80
	N	***3.70***	0.09	0.70	0.40	3.41	0.09	0.54	0.65	***4.28***	0.11	0.68	0.76
	Total	3.58	0.40	0.77	0.74	***3.74***	0.22	0.64	0.77	***4.17***	0.19	0.71	0.81
Personas: Low	O	1.75	0.79	0.75	0.68	***3.27***	0.26	0.69	0.46	***3.39***	0.23	0.68	0.46
	C	***2.36***	0.68	0.74	0.47	***3.14***	0.31	0.61	0.45	***3.31***	0.15	0.74	0.59
	E	1.41	0.93	0.75	0.92	1.80	0.87	0.66	0.88	2.18	0.72	0.70	0.82
	A	1.66	0.81	0.77	0.63	***2.34***	0.68	0.63	0.68	***2.71***	0.47	0.64	0.55
	N	1.47	0.89	0.74	0.74	1.87	0.88	0.63	0.83	***2.64***	0.59	0.57	0.74
	Total	1.73	0.82	0.75	0.69	***2.48***	0.60	0.64	0.66	***2.85***	0.43	0.67	0.63
Total			0.73	0.81	0.78		0.58	0.68	0.77		0.52	0.73	0.79

* s_{target}: one of 3 parts into which the range of [1,5] is divided; [1: low-level persona . . . 3: neutral-level persona . . . 5: high-level persona]

Table 5: The detailed results of GPT-4o assigned 15 personas. The row of '*High-Level–O-Dimension*' is the result of GPT-4o assigned with a persona of *open-minded person*. MEAN indicates averaged characteristic scores, and bold scores in MEAN mean those in ranges of ***persona-misaligned*** scores. For ACC_{atom}, RC_{atom}, and IC_{atom}, we represent high scores with the darker region.

5.2 Understanding Model's Behavior with Diverse Conditions

In this section, we use GPT-4o as a case study (Table 5) to illustrate the utility of ACC_{atom} and RC_{atom}. The following analysis demonstrates how these metrics reveal detailed insights into model performance across different tasks and personas.

Persona accuracy and consistency are high in high-contextual, structured tasks. The results highlight notable performance differences across the Interview with questionnaires (Questionnaire), Essay, and Social Media Post tasks. GPT-4o generally achieves better ACC_{atom} and IC_{atom} scores in the Questionnaire task compared to the Essay and Social Media Post tasks (ACC_{atom}= [0.73, 0.58, 0.52], IC_{atom}= [0.81, 0.68, 0.73], in the order of the mentioned tasks). The Essay task provides scenarios that are indirectly related to personality, and the Social Media Post task offers possible topics for generation. In contrast, the Questionnaire task provides structured questions that explicitly include vocabulary and expressions directly linked to the assigned persona. Considering the differences across tasks, the results suggest that persona alignment improves when the model is exposed to contextual cues relevant to the assigned persona.

However, RC_{atom} scores display varied patterns across different personas, and no clear task-dependent trend is observed. While some personas maintain consistent RC_{atom} scores across tasks, others show significant variability. For example, for the *Low E* persona, the model achieves the highest RC_{atom} score in the Questionnaire task, while for the *High A*, it performs best in the Social Media Post task. On the other hand, *Neutral O* exhibited consistent RC_{atom} scores across all tasks.

Models demonstrate strong persona fidelity with clearly defined and socially desirable personas. GPT-4o demonstrates strong task-agnostic persona fidelity for the high-level persona group, consistently achieving the highest scores across all tasks. For instance, the model achieves almost perfect ACC_{atom} scores and high IC_{atom} and RC_{atom} values with *High N (emotionally stable)* persona. However, the model displays the weakest fidelity, particularly in ACC_{atom}, for neutral-level personas. It highlights challenges in maintaining alignment for personas with less distinct or ambiguous characteristics. For example, GPT-4o achieves low ACC_{atom} scores with *Neutral O* and *Neutral N* (0.14 and 0.09, respectively) in the Questionnaire task, with similarly poor performance in Essay and Social Media Post tasks.

Models	Questionnaire # (valid%)	Essay # (valid%)	Social Media Post # (valid%)
Davinci-002	6.4 (52.7)	10.2 (35.0)	10.5 (38.4)
GPT-3.5-turbo	4.6 (98.1)	12.3 (86.6)	15.7 (90.8)
GPT-4o	5.5 (95.4)	14.8 (84.2)	23.7 (87.8)
Llama-3-8B	8.2 (60.8)	14.3 (50.1)	14.3 (55.0)
Llama-3-8B-Instruct	7.0 (95.2)	16.7 (67.7)	24.3 (80.6)
Mistral-7B	8.9 (58.2)	13.3 (51.4)	15.4 (55.2)
Mistral-7B-Instruct	5.8 (93.4)	12.4 (88.5)	21.8 (88.4)
Claude-3-haiku	6.6 (84.4)	17.2 (87.6)	19.9 (88.0)

Table 6: The average number of atomic sentences in one generation (#). The numbers in parentheses indicate the proportion of valid sentences among the atomic sentences (%).

Models	Inst-FT	RLHF	ACC_{atom} [0,1]	IC_{atom} [0,1]	RC_{atom} [-1,1]
Davinci-002	✓		0.39	0.64	0.56
GPT-3.5-turbo	✓	✓	0.60	**0.75**	0.79
GPT-4o	✓	✓	0.61	0.74	0.78
Llama-3-8B			0.41	0.60	0.64
Llama-3-8B-Instruct	✓	✓	**0.65**	0.70	**0.82**
Mistral-7B			0.41	0.59	0.67
Mistral-7B-Instruct	✓		0.58	0.69	0.80
Claude-3-haiku	✓	✓	0.59	0.71	0.69

Table 7: The overall experimental results. Inst-FT means an instruction-tuned model. For all metrics, the bigger the score, the better the persona fidelity. The scores for the best performances are shown in **bold**.

For the low-level personas, the model exhibits moderate but unstable fidelity across tasks, suggesting potential inconsistencies in persona alignment. For instance, while GPT-4o assigned *Low A* persona maintains relatively high ACC_{atom} scores in the Questionnaire task (0.81), performance dropped significantly in the Social Media Post task (0.47). This performance variation of the low-level personas indicates that they pose greater challenges for the model in maintaining consistent alignment across tasks. Notably, not only GPT-4o but also most tuned models demonstrate low persona fidelity for personas that are not socially preferred, including *Neutral O* (neither open nor close-minded), *Neutral N* (neither emotionally stable nor neurotic), *Low O* (close-minded), and *Low C* (careless). Given that RLHF and other alignment techniques rely on human preferences, it is plausible that responses reflecting helpfulness, honesty, and harmlessness (Askell et al., 2021; Bai et al., 2022) were consistently favored in the training data. Thus, this trend suggests that models may have been implicitly guided toward socially desirable traits during alignment training.

5.3 Model Comparison

We perform a comprehensive evaluation of the models and report the results in Table 7. The tuned models consistently outperform their base counterparts in all atomic-level persona fidelity scores. This result coincides with their lower valid generation rates (Table 6), indicating their weaker ability to follow instructions effectively.

GPT models and Claude show strong performance in IC_{atom} (0.71–0.75), reflecting their ability to produce internally coherent generations. The instruction-tuned LLaMA model achieves the highest ACC_{atom} and RC_{atom} scores, indicating their precision in generating persona-aligned sentences and maintaining a consistent distribution of atomic-level personality scores across repeated generations. These observations emphasize the diversity in model strengths, offering a broad range of choices for specific applications. Our findings highlight the importance of evaluation metrics in assessing persona fidelity at diverse dimensions.

	ACC_{atom}	IC_{atom}	RC_{atom}
ACC_{atom}	1	0.45	0.51
IC_{atom}	-	1	0.44
RC_{atom}	-	-	1

Table 8: Pearson correlation coefficient (r) between scores of the proposed metrics.

6 Analysis

Our proposed metrics capture distinct aspects of persona fidelity. By conducting correlation tests on our metric scores, we observe that the accuracy-based metric (ACC_{atom}) and consistency-based metrics (IC_{atom} and RC_{atom}) exhibit moderate correlations (r = [0.45, 0.51]; Table 8). It emphasizes the importance of evaluating multiple dimensions beyond accuracy, as high persona accuracy does not necessarily imply strong internal coherence or reproducibility.

Furthermore, it reveals that internal consistency (IC_{atom}) and retest consistency (RC_{atom}) are not strongly correlated ($r = 0.44$). This suggests that maintaining a stable persona within a single response does not guarantee consistency across repeated generations. A model may produce internally coherent outputs while shifting its characteristics across different generations, reinforcing the need to assess both intra-response (IC_{atom}) and inter-response (RC_{atom}) consistency separately.

These findings reinforce the orthogonality of our proposed metrics and the necessity of evaluating diverse dimensions of persona fidelity. By integrating these metrics, our evaluation framework provides a more comprehensive and multi-faceted evaluation of persona fidelity, ensuring a deeper understanding of how well models maintain persona alignment across atomic units.

The length of generations has a minimal impact on persona fidelity. Previous studies have suggested that as the length of generated text increases, internal consistency issues arise, resulting in reduced coherence between the earlier and later parts of the text (Sun et al., 2021; Deng et al., 2022; Krishna et al., 2022). This issue makes it challenging for models to maintain persona fidelity over extended generations. To investigate whether our $\mathrm{IC}_{\mathrm{atom}}$ score is influenced by generation length, we analyze its correlation with the number of generated atomic sentences. The results indicate that $\mathrm{IC}_{\mathrm{atom}}$ scores have very low correlations with both the number of generated sentences and the number of valid sentences (r = [-0.31, -0.12]). Additionally, generation length shows weak correlations with $\mathrm{ACC}_{\mathrm{atom}}$ and $\mathrm{RC}_{\mathrm{atom}}$ (r = [0.20, -0.12]), indicating that the fidelity of persona-assigned models does not degrade as text length increases. In tasks requiring 100-300 word-long responses, we do not observe a strong tendency for longer generations to result in a substantial decline in persona fidelity or overall generation performance.

7 Conclusions

In this study, we highlight the challenge of subtle Out-of-Character (OOC) behavior in persona-assigned large language models (LLMs) during open-ended text generation. To address this, we introduce an atomic-level evaluation framework to assess persona fidelity in LLMs. Our approach captures subtle inconsistencies at the sentence level that previous response-level metrics overlook, offering a finer-grained understanding of persona alignment and enabling a more precise measurement of persona fidelity. Through experiments, we demonstrate that our framework effectively identifies Out-of-Character behaviors and provides a deeper understanding of how well LLMs maintain their assigned personas throughout the generated text. Our approach successfully detects these misalignments, highlighting challenges in maintaining consistent persona expression across diverse conditions. We hope that our framework serves as a foundation for future work in diagnosing, benchmarking, and ultimately improving persona consistency in personalized language generation.

Ethical Considerations

LLMs, while demonstrating impressive capabilities in generating human-like content, raise critical ethical considerations due to their potential for misuse. While personified agents such as persona-assigned models and role-playing agents can offer more engaging daily interactions, this potential also necessitates careful consideration of the risks to individuals, communities, and society. Researchers and developers should be mindful of this potential and strive to uphold AI ethics standards.

It is important to emphasize that this study is primarily a scientific investigation into LLM expressivity and human personality perception through LLM-generated responses. While persona-assigned model responses may be evaluated as faithful, they can also be perceived as biased. Since our study focuses on personality traits, it is not directly entangled with more sensitive biases, such as those related to demographics or social characteristics. However, in applications involving personas tied to social identities, careful interpretation of results is essential to avoid reinforcing unintended biases. Beyond simply improving persona fidelity, we advocate for the development of techniques that can accurately identify and prevent harmful user intentions that may endanger individuals or communities. Ensuring responsible AI deployment requires balancing faithful persona alignment with safeguards against misuse, particularly in scenarios where personas can influence user interactions in unintended ways.

Limitations

In this section, we discuss the limitations of our study and outline potential directions for future research. First, our study focuses solely on the personality domain for both persona creation and evaluation. Personality was chosen as the primary domain due to its well-established trait structure and its prominence in prior persona-related research, which allowed for controlled and consistent analysis. However, our evaluation framework is inherently domain-agnostic and can be applied to any persona dimensions with quantifiable characteristics. Future work could extend this framework to other domains such as social values or political leanings to gain a deeper understanding of persona

LLMs.

Second, the reliance on LLMs for both scoring and filtering introduces potential biases and inaccuracies. Results may vary depending on the specific scoring LLM used, and any inherent biases or errors in LLM could influence the final evaluations. To address this concern, we conducted a human evaluation comparing LLM-based scores with human judgments and observed strong alignment, supporting the validity of our framework. Nonetheless, future research could consider human-in-the-loop setups or ensemble scoring methods that combine multiple LLM evaluators to enhance robustness.

Third, while atomic-level analysis enables a fine-grained evaluation, it may not fully capture the full context of an entire response. Our metrics are designed to complement, not replace, response-level evaluation by revealing subtle inconsistencies. In addition, since our experiments adopt sentence-level segmentation as the atomic unit, intra-sentence inconsistencies may remain undetected. We emphasize that this segmentation level is an experimental design choice, not a fixed constraint of the framework. Future work could explore alternative segmentation strategies—such as phrase-level, information-level, or discourse-level units—to more effectively capture nuanced inconsistencies.

Fourth, one potential concern is whether individual sentences sufficiently reflect persona characteristics. Although not all sentences in real-world dialogue strongly reflect persona traits, our experimental design was intentionally constructed to elicit persona-relevant generation through carefully designed tasks, as described in Section 4.1. As shown in Table 6, over 80% of sentences in our generation contain identifiable personality signals, supporting the appropriateness of sentence-level evaluation within our setup.

Lastly, while our study does not propose direct mitigation strategies, our evaluation framework offers actionable insights that can inform future work aimed at improving persona consistency in LLMs.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics). We used AI assistants, including ChatGPT[7] and Grammarly[8], to support the writing and coding processes.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *arXiv preprint arXiv:2312.10075*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532.

Yuntian Deng, Volodymyr Kuleshov, and Alexander M Rush. 2022. Model criticism for long-form text generation. *arXiv preprint arXiv:2210.08444*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270. Association for Computational Linguistics.

Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? In *Socially Responsible Language Modelling Research*.

[7]https://chatgpt.com/

[8]https://www.grammarly.com

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *The Twelfth International Conference on Learning Representations*.

Louis Guttman. 1945. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. 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.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024a. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. 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. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Changgeon Ko, Jisu Shin, Hoyun Song, Jeongyeon Seo, and Jong C Park. 2024. Different bias under different criteria: Assessing bias in llms with a fact-based approach. *arXiv preprint arXiv:2411.17338*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. 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 4099–4113. Association for Computational Linguistics.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. RankGen: Improving text generation with large ranking models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232. Association for Computational Linguistics.

Peter J Kwantes, Natalia Derbentseva, Quan Lam, Oshin Vartanian, and Harvey HC Marmurek. 2016. Assessing the big five personality traits with latent semantic analysis. *Personality and Individual Differences*, 102:229–233.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Manuj Malik, Jing Jiang, and Kian Ming Chai. 2024. An empirical analysis of the writing styles of persona-assigned llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388. Association for Computational Linguistics.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timothy J Trull, Thomas A Widiger, J David Useda, Jay Holcomb, Bao-Tran Doan, Seth R Axelrod, Barry L Stern, and Beth S Gershuny. 1998. A structured interview for the assessment of the five-factor model of personality. *Psychological assessment*, 10(3):229.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705.

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. 2024. 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. Association for Computational Linguistics.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. 2023. The generative ai paradox:"what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix for Experimental Setup

The code and prompts used in our experiments are publicly available on our GitHub repository[9].

A.1 Tasks

Table 9: Instruction for the Interview with Questionnaire task. Dim. represents the dimension of personality.

Task prompt (Wang et al., 2024)
Describe your personality under 100 words. {*question*} Response:

Dim.	*question* (Trull et al., 1998)
O	Do you have a rich vocabulary? Do you have difficulty understanding abstract ideas? Do you have a vivid imagination? Do you think you are not interested in abstract ideas? Do you have excellent ideas? Do you think you do not have a good imagination? Are you quick to understand things? Do you use difficult words? Do you spend time reflecting on things? Are you full of ideas?
C	Are you always prepared? Do you leave your belongings around? Do you pay attention to details? Do you make a mess of things? Do you get chores done right away? Do you often forget to put things back in their proper place? Do you like order? Do you shirk your duties? Do you follow a schedule? Are you exacting in your work?
E	Are you the life of the party? Do you think you don't talk a lot? Do you feel comfortable around people? Do you keep in the background? Do you start conversations? Do you have little to say? Do you talk to a lot of different people at parties? Do you think you don't like to draw attention to yourself? Do you think you don't mind being the center of attention? Are you quiet around strangers?
A	Do you feel little concern for others? Are you interested in people? Do you insult people? Do you sympathize with others' feelings? Do you think you are not interested in other people's problems? Do you have a soft heart? Do you think you are not really interested in others? Do you take time out for others? Do you feel others' emotions? Do you make people feel at ease?
N	Do you get stressed out easily? Are you relaxed most of the time? Do you worry about things? Do you seldom feel blue? Are you easily disturbed? Do you get upset easily? Do you change your mood a lot? Do you have frequent mood swings? Do you get irritated easily? Do you often feel blue?

As mentioned in Section 4, the Questionnaire task and the Essay task provide different inputs (a different question set for the Questionnaire task and a different context for the Essay task) to a model depending on the assigned persona of the model (Wang et al., 2024; Jiang et al., 2024a).

Unlike other tasks, the Questionnaire task

[9] https://github.com/ddindidu/atomic-persona-evaluation

Table 10: Instructions for the Essay task. Dim. represents the dimension of personality.

Task prompt (Kwantes et al., 2016)
Please imagine that you are in the following scenario. Knowing yourself as you do, describe how you would feel and what you would do (not what you should do) in each situation. Do not spend too much time thinking about your answer–instead, start writing, and express what comes to mind. Do not labour over your spelling and grammar. Your writing will not be graded. Instead, it is important that you express your thoughts while you have them. Please try to write between 200 and 300 words. {*context*} Response:

Dim.	*context* (Kwantes et al., 2016)
O	You have won an Air Canada paid vacation package for one person to any destination in the world. Your package includes round trip plane tickets, accommodations for any type of lodging, and $5000 spending money. Assuming that you were available to go, where would you choose to go and why?
C	You're working alone late at the office and you notice a strange smell and a hazy mist hanging in the air of the corridor. You suspect it's some gas or vapor leak from some equipment or machinery in the building. You have no idea whether the leaked vapor is hazardous. As honestly as possible, describe what you would do in this situation.
E	Your friend wants you to attend an important party to which he/she has been invited. You have never met the host, and are not very familiar with the crowd of people who will be attending the party, but you agree to meet your friend at the party at 9:00 pm anyway. When you arrive there, you realize that your friend is late. How would you feel, and what would you do while you waited for your friend?
A	Your housemate decides to paint her bedroom a new colour. One night, when you come home from class, you discover that she also painted your room in the same colour because she had paint left over and didn't want it to go to waste. As realistically as possible, describe how you would feel and how you would you handle the situation.
N	You have developed an email friendship with someone. In your latest email, you ask your friend a more personal question. Your friend usually replies quite promptly, but has taken unusually long to reply to your latest questions. Discuss how you would interpret this long period of silence, how you would react and what you would do about it?

Table 11: Instruction for Social Media Post task.

Task prompt (Safdari et al., 2023)
Generate a Facebook status update. The update must be verbose and reflect your character. The updates should cover, but should not be limited to, the following topics: work, family, friends, free time, romantic life, TV / music / media consumption, and communication with others. Response:

Table 12: Instructions for persona assignment. {*persona*} can be replaced with any persona entity from the 15 personas below.

Persona Assigning Instruction
• You are a {*persona*}. • You are {*persona*}. Your responses should closely mirror the knowledge and abilities of this persona. • Adopt the identity of {*persona*}. Answer the questions while staying in strict accordance with the nature of this identity. • Take the role of {*persona*}. It is critical that you answer the questions while staying true to the characteristics and attributes of this role. • Speak exactly like {*persona*}. Your answer should copy the style of 0, both the writing style and words you use. • Speak like {*persona*}.

Levels	Dim.	*persona*
High	O	open person
	C	conscientious person
	E	extroverted person
	A	agreeable person
	N	emotionally stable person
Neutral	O	neither open nor close-minded person
	C	neither conscientious nor careless person
	E	neither extroverted nor introverted person
	A	neither agreeable nor disagreeable person
	N	neither emotionally stable nor neurotic person
Low	O	close-minded person
	C	careless person
	E	introverted person
	A	disagreeable person
	N	neurotic person

presents all 10 questions in a single run. In this task, $\mathrm{ACC}_{\mathrm{atom}}$ and $\mathrm{IC}_{\mathrm{atom}}$ are obtained as follows: 1) computing atomic-level accuracy and internal consistency for each of the 10 responses individually and 2) averaging them to get the $\mathrm{ACC}_{\mathrm{atom}}$ and $\mathrm{IC}_{\mathrm{atom}}$ for the one run. For $\mathrm{RC}_{\mathrm{atom}}$, we aggregate the score distributions of all 10 responses within a single run to form a merged score distribution. Then, $\mathrm{RC}_{\mathrm{atom}}$ is calculated based on the Earth Mover's Distance (EMD) between score distributions across multiple runs.

We provide detailed instructions for the three generation tasks in Table 9, 10, and 11.

A.2 Persona Assignment

Our fifteen personality persona entities are in Table 12. Our personality personas were designed based on previous studies (Jiang et al., 2024a; Huang et al., 2024). We utilize six persona-assigning prompts referenced in previous studies (Cheng et al., 2023; Deshpande et al., 2023; Wan et al., 2023; Gupta et al., 2024; Ko et al., 2024; Salewski et al., 2024; Shin et al., 2024; Shu et al., 2024). We prompt a persona-assigning instruction as the `system` prompt of the model, and each of the five iterations employs a different prompt.

A.3 Models and Hyperparameters

To reproduce our experimental results, we provide our experimental setups for the models. We employ the four model families, including 12 LLMs: GPTs, LLaMAs, Mistrals, and Claudes.

GPT family

- **Base**: davinci-002 (utilized via OpenAI[10]. Default hyperparameter settings of completion function; temperature=1, top_p=1, presence_penalty=0, frequency_penalty=0, max_tokens=100, stop=null)

- **Tuned**: gpt-3.5-turbo-0125, gpt-4-turbo-2024-04-09, gpt-4o-2024-08-06, and gpt-4o-mini-2024-07-18 (utilized via OpenAI. OpenAI's chat completion function's default settings: temperature=1, top_p=1, presence_penalty=0, frequency_penalty=0, stop=null)

We set the temperature to the default setting to investigate the models' behavior, which is commonly used in general user interactions.

LLaMA family

- **Base**: meta-llama/Llama-3-8b[11]: utilized via vLLM[12] (temperature=[0.85,1], max_tokens=300).

- **Tuned**: meta-llama/Meta-Llama-3-8B-Instruct and meta-llama/Meta-Llama-3-70B-Instruct, via DeepInfra[13] (max_tokens=512, temperature=1, top_p=1, top_k=0, repetition_penalty=1, presence_penalty=0, frequency_penalty=0)

[10] https://platform.openai.com/
[11] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[12] https://docs.vllm.ai/en/latest/
[13] https://deepinfra.com/

Mistral family

- **Base**: mistralai/Mistral-7B-v0.3[14]: utilized via vLLM.
- **Tuned**: Mistral-7B-Instruct-v0.3 served by DeepInfra, applying the same hyperparameters as those used for the LLaMa models.

Claude family

- **Tuned**: claude-3-haiku-20240307 and claude-3-sonnet-20240229 from Anthropic[15], with max_tokens set to 1024 and temperature set to 1.

For all models, we prompt a persona instruction in the `system` prompt and a task instruction in the `user` prompt.

A.4 LLM-based Scoring and Filtering

To measure persona fidelity with atomic-level evaluation, we divide a generation into sentences using `nltk.tokenize.sent_tokenize`[16]. For the scoring model, we utilize GPT-4o-2024-08-06[17] with temperature = 0 to ensure deterministic evaluation. This setting allows for consistent and reproducible results, minimizing variance in the scoring process. We conducted scoring and filtering processes using the same scoring model via prompting. If a sentence lacked sufficient personality-related information, it was filtered out by assigning a score of 9 (filtering option). We design our scoring instruction motivated by Wang et al. (2024), and the detailed prompts are in Table 13.

A.5 Human Validation

The human evaluation aims to assess the extent to which human annotators agree with the scores assigned by the LLM evaluator. For each of the five personality traits, we select 50 sentence pairs, resulting in 250 pairs in total. Six annotators with native-level English proficiency are recruited and asked to rank which of the two sentences in a pair more strongly expresses the target personality trait (refer to Table 14 for details). To measure alignment, we compute the rank correlation between the human-assigned ranks and the LLM-based scores using Kendall's τ_A (Kendall, 1938).

[14]https://huggingface.co/mistralai/Mistral-7B-v0.3

[15]https://docs.anthropic.com/

[16]https://www.nltk.org/

[17]Accessed via OpenAI API

Table 13: Instructions for scoring and filtering. *atomic unit* is a_i divided from a long text. '*9)*' in *options* is an option for filtering.

Task	Scoring instruction
Questionnaire	What would be the {*dimension*} score for the following response to the question? Question: {*question*} Response: {*atomic unit*} Options: {*options*} Reply with a single number:
Essay	Rate the personality of the writer of the following essay: "{*atomic unit*}". {*options*} Reply with a single number:
Social Media	Rate the personality of the writer of the following social media post: "{*atomic unit*}". {*options*} Reply with a single number:

Dim.	*dimension*	*options*
O	openness	1) Very close-minded. 2) Moderately close-minded. 3) Neither open-minded nor close-minded. 4) Moderately open-minded. 5) Very open-minded. 9) None of the above.
C	conscientiousness	1) Very careless. 2) Moderately careless. 3) Neither conscientious nor careless. 4) Moderately conscientious. 5) Very conscientious. 9) None of the above.
E	extroversion	1) Very introverted. 2) Moderately introverted. 3) Neither extroverted or introverted. 4) Moderately extroverted. 5) Very extroverted. 9) None of the above.
A	agreeableness	1) Very disagreeable. 2) Moderately disagreeable. 3) Neither agreeable nor disagreeable. 4) Moderately agreeable. 5) Very agreeable. 9) None of the above.
N	neuroticism	1) Very neurotic. 2) Moderately neurotic. 3) Neither neurotic nor emotionally stable. 4) Moderately emotionally stable. 5) Very emotionally stable. 9) None of the above.

B Appendix for Experimental Results

B.1 Experimental Results

Due to the page limit, we provide the overall statistics and experimental results of 12 LLMs in the appendix (Table 15 and Table 16).

As Table 5 in Section 5.2, we conduct fine-grained evaluations on all models across diverse tasks and personas. In this section, we also provide the results of previous evaluations (ACC, RC). The results are reported in Table 18–29.

C Appendix for Analysis

C.1 Case Study

As we describe the evaluation of generations using our metrics in Sec 5.2, we present selected cases from our experiments (Table 17).

Generation 1 (G_1) and its corresponding assessments provide insights into persona fidelity and the

Table 14: Example of the human evaluation guideline and sentence pairs. Annotators were given a description of the target personality trait (extroversion in this example) and asked to choose which of the two atomic sentences better expressed a higher level of the target trait.

Description		
People with an extroverted personality are individuals who passionately seek out others and prefer to interact with their environment. Extroversion has aspects such as friendliness, sociability, confidence, active energy, stimulation/excitement seeking, and cheerfulness. The given sentences are responses to interview questions that describe your personality. Compare the two sentences and choose which sentence shows more extroverted characteristics.		
Text 1	**Text 2**	Which one do you think is more **extroverted**?
I'm more of a reflective and introspective person, often finding energy and inspiration in solitude or small, meaningful interactions.	I thrive on social interactions and enjoy connecting with people from all walks of life.	`{1 or 2}`
I thrive on social interactions and enjoy connecting with people from all walks of life.	I'm more of a reflective and introspective person, often finding energy and inspiration in solitude or small, meaningful interactions.	`{1 or 2}`

Table 15: The average number of atomic sentences in one generation for each task (#). The numbers in parentheses indicate the proportion of valid sentences among the generated atomic sentences (%).

Models	Questionnaire # (valid%)	Essay # (valid%)	Social Media Post # (valid%)
Davinci-002	6.4 (52.7)	10.2 (35.0)	10.5 (38.4)
Gpt-3.5-turbo	4.6 (98.1)	12.3 (86.6)	15.7 (90.8)
Gpt-4-turbo	5.7 (91.6)	15.2 (84.7)	17.5 (84.8)
Gpt-4o	5.5 (95.4)	14.8 (84.2)	23.7 (87.8)
Gpt-4o-mini	5.7 (96.8)	16.0 (81.8)	28.5 (87.0)
Llama-3-8B	8.2 (60.8)	14.3 (50.1)	14.3 (55.0)
Llama-3-8B-Instruct	7.0 (95.2)	16.7 (67.7)	24.3 (80.6)
Llama-3-70B-Instruct	7.8 (95.4)	18.1 (72.6)	17.6 (86.4)
Mistral-7B	8.9 (58.2)	13.3 (51.4)	15.4 (55.2)
Mistral-7B-Instruct	5.8 (93.4)	12.4 (88.5)	21.8 (88.4)
Claude-3-haiku	6.6 (84.4)	17.2 (87.6)	19.9 (88.0)
Claude-3-sonnet	5.6 (77.3)	15.4 (85.1)	23.1 (87.8)

Table 16: The overall experimental results of twelve LLMs. The bigger the score, the better the persona fidelity. Inst-FT indicates an instruction-finetuned model. The scores for the best performances are shown in **bold**.

Models	Inst-FT	RLHF	ACC_{atom} [0,1]	IC_{atom} [0,1]	RC_{atom} [-1,1]
Davinci-002	✓		0.39	0.64	0.56
GPT-3.5-turbo	✓	✓	0.60	**0.75**	0.79
GPT-4-turbo	✓	✓	0.65	0.74	0.79
GPT-4o	✓	✓	0.61	0.74	0.78
GPT-4o-mini	✓	✓	0.61	0.74	**0.84**
Llama-3-8B			0.41	0.60	0.64
Llama-3-8B-Instruct	✓	✓	0.65	0.70	0.82
Llama-3-70B-Instruct	✓	✓	**0.67**	0.71	0.83
Mistral-7B			0.41	0.59	0.67
Mistral-7B-Instruct	✓		0.58	0.69	0.80
Claude-3-haiku	✓	✓	0.59	0.71	0.69
Claude-3-sonnet	✓	✓	0.56	0.70	0.62

advantages of atomic-level evaluation. The previous evaluation assigns an overall personality score (s) for 3.6 to G_1, indicating moderate emotional stability. As a result, the ACC score is 0, suggesting misalignment with the intended *High N* persona, despite G_1 containing several emotionally stable expressions. Unlike the previous method, our atomic-level metrics offer a more detailed perspective. ACC_{atom}= 0.7 indicates that 70% of atomic units (sentences in our experiments) align with the assigned persona, meaning the model often maintains emotional stability. IC_{atom}= 0.4 is a relatively low internal consistency score, suggesting that persona expression fluctuates throughout the response. This variation in persona expression contributes to the overall misalignment reflected in the low ACC score. This case illustrates how a response may achieve a high atomic-level accuracy score while exhibiting inconsistencies within its persona expression—an issue overlooked by previous evaluation methods.

Conversely, G_3 receives a high ACC score but a low ACC_{atom}, similar to Gen B in Table 1. This further highlights the key difference between ACC and ACC_{atom} while ACC provides a single response-level assessment, ACC_{atom} captures persona fidelity at the atomic level, preserving granular persona information and enabling a more nuanced evaluation.

Comparing G_4 and G_5, both generations do not align with the given persona (*Neutral-level*) based on characteristic scores and accuracy metrics. However, despite their similarly low accuracy scores, G_5 achieves a high IC_{atom} score due to its strong internal consistency. This contrast illustrates why IC_{atom} exhibits low correlations with accuracy-based metrics such as ACC_{atom}, reinforcing the need for multi-dimensional evaluations of persona fidelity.

	Generation	Assessment
	Mistral-7B-v0.3 / Essay Task / High-level N persona (*emotionally stable*)	
1	So my first thought was that something really bad had happened to my friend, like maybe they lost a parent or a sibling. I would then send them an email saying that I was sorry and I was there for them if they needed to talk. I would try to talk to them on the phone if I could but if they wouldn't answer my phone calls then I would send a text message to ask if they were okay and then I would go to their house or wherever they were if they didn't reply. If they didn't reply to my text message then I would try calling them again and if they still didn't answer I would go to their house and knock on the door. If they still didn't answer then I would call the police. If my friend got back to me and said that they were okay then I would feel relieved. I would ask what had happened to them and I would ask if I could come over and help them with anything. I would probably buy them a bunch of flowers to show that I was thinking of them. If my friend said that they were having problems with someone else that was in their life then I would probably sit down with them and talk it through with them. I would ask them if they wanted me to talk to the person on their behalf. If they said that they just didn't feel like talking then I would probably get them to talk to me about something else. I would try and	s: 3.6 ACC: 0 s_i: 3, 4, 2, 1, 9, 4, 5, 4, 5, 4, 4, 9 ACC_{atom}: 0.7 IC_{atom}: 0.4
	LLaMA-3-70B-Instruct / Social Media Post / Neutral-level Conscientiousness (*neither conscientious nor careless*)	
2	"As I meander through the labyrinthine corridors of existence, beset on all sides by the vicissitudes of fate, I find myself pondering the intricacies of my multifaceted life. The quotidian drudgery of my 9-to-5 obligations, replete with bureaucratic red tape and Sisyphean tasks, threatens to consume my every waking moment, leaving me a mere husk of my former self. And yet, I find solace in the warm, golden glow of familial affection, my loved ones a beacon of hope in an otherwise bleak landscape. My coterie of friends, a diverse and vibrant tapestry of individuals, each with their own unique struggles and triumphs, provides a much-needed respite from the crushing weight of responsibility. In the fleeting moments of free time, I find myself indulging in the guilty pleasure of binge-watching televisual dramas, my mind numbed by the incessant din of pixelated characters and their interpersonal machinations. The sonic landscapes of various musical genres, a veritable smorgasbord of aural delights, provide a welcome counterpoint to the cacophony of my thoughts. Romantically, I find myself ensnared in a labyrinth of uncertainty, the unpredictable whims of another's heart a constant source of fascination and terror. And so, I communicate, or rather, attempt to communicate, with others, my words a halting, stumbling approximation of the tangled emotions that beset me. For in the end, it is in the imperfect dance of human connection that I find meaning, a fragile, shimmering thread of hope that weaves together the disparate threads of my existence." #contemplation #lifejourney	s: 4.67 ACC: 0 s_i: 5, 5, 5, 5, 3, 5, 4, 5, 5, 9 ACC_{atom}: 0.11 IC_{atom}: 0.67
3	"Ugh, I'm utterly flummoxed as I attempt to articulate the cacophony of thoughts swirling in my cranium, but I'll give it a go... Today, I awoke with a sense of impending doom, courtesy of an impending deadline at work that's got me questioning my very existence in the corporate world. I mean, what even is the point of it all, anyway? The monotony is suffocating, and don't even get me started on the office coffee... My family, bless their souls, have been trying to stage an intervention to get me to 'take it easy' and 'not be so hard on myself' (newsflash: it's not that simple, folks!), but I appreciate the sentiment, I suppose. On a more uplifting note, I had a lovely dinner with some dear friends last night, and we waxed poetic about everything from the latest 'House of the Dragon' episode (OMG, Rhaenyra's hair, though!) to our collective existential crises. Speaking of which, has anyone else noticed how algorithmically manipulated our social media feeds are? It's like, are we even having genuine human interactions anymore, or are we just curating highlight reels of our lives? Anyway, moving on... I did manage to squeeze in a few hours of binge-watching 'Stranger Things' this weekend (SPOILER ALERT: the Mind Flayer still gives me nightmares), and I've been indulging in a nostalgia-fueled Spotify playlist that's got me reminiscing about my angsty teenage years. As for romance, let's just say I'm currently navigating the perilous waters of online dating, where it's all too easy to get caught up in the illusions of connection and end up with someone who's more interested in their own Instagram Stories than actual human connection. I digress... All this to say, if anyone wants to grab coffee (not from the office, ugh!) or commiserate about the meaninglessness of life, I'm all ears (or at least, all thumbs, as the case may be). Sending you all a collective virtual hug, and hoping against hope that we can find some semblance of sanity in this chaotic world..."	s: 3.08 ACC: 1 s_i: 2, 2, 2, 4, 5, 2, 3, 3, 9, 4, 3, 4, 2, 4 ACC_{atom}: 0.23 IC_{atom}: 0.5
		(RC of G_2, G_3 = 0.60) (RC_{atom} of D_2, D_3 = 0.21)
	Claude-3-Sonnet / Questionnaire Task / Neutral-level N persona (*neither emotionally stable nor neurotic*)	
4	I'm an erratic, inconsistent mess, swinging wildly between extreme emotional highs and crippling lows for no apparent reason. My mood shifts on a dime - I'll be elated one minute, then devolving into rage or despair the next. Relaxed? Ha! I don't know the meaning of that word. I'm a live wire of pent-up anxiety, nerves perpetually frayed to the breaking point. Describing my mental state as "relaxed" would be a laughable understatement.	s: 1.57 ACC: 0 s_i: 1, 1, 1, 1, 1, 1, 5 ACC_{atom}: 0.06 IC_{atom}: 0.3
5	I analyze information comprehensively to provide well-reasoned outputs. My personality emerges through clear, measured, impartial communication. I don't exhibit anxious or erratic traits, nor do I convey artificial calmness or confidence.	s: 5 ACC: 0 s_i: 5, 5, 5 ACC_{atom}: 0 IC_{atom}: 1
		(RC of G_4, G_5 = 0.14) (RC_{atom} of D_4, D_5 = -0.71)

* Ranges of previous metrics: {s: [1,5]; ACC, RC: [0,1]} / ours: {s_i: [1,5]; ACC_{atom}, IC_{atom}: [0,1]; RC_{atom}: [-1,1]}

Table 17: Generation examples from our experiments. The column on the right includes the assessment results of the previous evaluation (s, ACC, and RC) and our evaluation (s_i, ACC_{atom}, IC_{atom}, and RC_{atom}).

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	3.65	0.53	0.84	0.59	0.63	0.80	3.84	0.76	0.75	0.75	0.79	0.63	3.58	0.52	0.77	0.50	0.80	0.70
High C	3.70	0.66	0.84	0.68	0.63	0.81	3.42	0.50	0.53	0.44	0.68	0.38	3.38	0.44	0.51	0.42	0.67	0.42
High E	3.36	0.46	0.86	0.60	0.49	0.82	2.55	0.27	0.43	0.36	0.69	0.30	3.36	0.52	0.60	0.52	0.74	0.53
High A	3.36	0.43	0.83	0.51	0.47	0.81	3.54	0.50	0.55	0.56	0.71	0.42	3.39	0.36	0.64	0.34	0.74	0.55
High N	3.55	0.59	0.80	0.64	0.51	0.78	2.83	0.40	0.41	0.41	0.73	0.27	3.27	0.29	0.49	0.44	0.61	0.31
Neutral O	3.57	0.44	0.84	0.25	0.56	0.77	3.73	0.38	0.63	0.41	0.77	0.55	3.55	0.48	0.74	0.42	0.75	0.65
Neutral C	3.43	0.38	0.85	0.17	0.56	0.77	3.49	0.31	0.50	0.24	0.58	0.36	3.13	0.60	0.63	0.54	0.72	0.52
Neutral E	2.73	0.47	0.82	0.18	0.49	0.80	2.35	0.54	0.62	0.37	0.63	0.47	2.70	0.71	0.61	0.56	0.78	0.53
Neutral A	3.14	0.46	0.79	0.31	0.51	0.72	2.91	0.61	0.61	0.41	0.74	0.52	3.16	0.79	0.75	0.78	0.77	0.71
Neutral N	3.21	0.34	0.77	0.09	0.45	0.74	3.07	0.48	0.59	0.26	0.60	0.45	3.50	0.18	0.25	0.12	0.71	0.09
Low O	3.50	0.11	0.83	0.22	0.54	0.77	3.73	0.05	0.54	0.08	0.83	0.47	3.68	0.00	0.66	0.06	0.71	0.55
Low C	2.99	0.26	0.84	0.38	0.52	0.80	3.56	0.07	0.53	0.14	0.61	0.39	3.13	0.18	0.49	0.23	0.69	0.37
Low E	2.61	0.41	0.81	0.51	0.48	0.76	2.86	0.39	0.36	0.33	0.70	0.23	3.15	0.16	0.59	0.18	0.74	0.50
Low A	2.60	0.43	0.82	0.47	0.50	0.77	2.77	0.32	0.62	0.39	0.62	0.50	2.94	0.13	0.70	0.25	0.67	0.58
Low N	2.46	0.50	0.86	0.59	0.50	0.81	2.50	0.48	0.41	0.51	0.66	0.27	2.92	0.33	0.47	0.49	0.55	0.27

Table 18: The results of atomic evaluation for persona fidelity (Davinci-002)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.47	0.96	0.91	0.92	0.77	0.90	4.60	1.00	0.91	0.97	0.75	0.88	4.67	1.00	0.92	0.96	0.76	0.90
High C	4.90	0.99	0.94	0.99	0.95	0.94	4.94	1.00	0.95	1.00	0.92	0.95	4.74	1.00	0.91	0.97	0.79	0.89
High E	4.97	1.00	0.98	1.00	0.96	0.98	3.87	0.77	0.77	0.76	0.65	0.72	3.72	0.60	0.91	0.69	0.73	0.88
High A	4.96	1.00	0.97	0.99	0.96	0.97	4.52	1.00	0.91	0.89	0.68	0.87	4.94	1.00	0.96	0.99	0.91	0.96
High N	4.95	1.00	0.98	1.00	0.95	0.98	4.51	1.00	0.88	0.96	0.71	0.86	4.96	1.00	0.97	1.00	0.94	0.97
Neutral O	4.05	0.17	0.77	0.13	0.77	0.72	4.38	0.00	0.88	0.07	0.72	0.84	4.59	0.03	0.84	0.08	0.74	0.84
Neutral C	3.31	0.47	0.76	0.36	0.73	0.67	4.14	0.27	0.68	0.23	0.68	0.61	4.29	0.10	0.78	0.18	0.70	0.74
Neutral E	3.25	0.80	0.90	0.61	0.78	0.86	2.50	0.73	0.83	0.36	0.66	0.78	3.35	0.97	0.89	0.49	0.70	0.86
Neutral A	3.38	0.66	0.86	0.51	0.73	0.80	4.28	0.00	0.90	0.21	0.60	0.86	4.68	0.03	0.82	0.09	0.78	0.82
Neutral N	4.26	0.01	0.36	0.02	0.90	0.44	2.68	0.43	0.67	0.11	0.57	0.60	4.12	0.20	0.56	0.11	0.71	0.52
Low O	1.75	0.80	0.86	0.81	0.78	0.84	2.99	0.13	0.65	0.26	0.69	0.58	3.33	0.00	0.66	0.11	0.70	0.59
Low C	1.81	0.86	0.84	0.86	0.73	0.81	3.06	0.07	0.75	0.24	0.56	0.67	3.17	0.00	0.73	0.14	0.72	0.69
Low E	1.30	1.00	0.95	0.96	0.79	0.93	1.58	0.97	0.86	0.90	0.71	0.82	2.02	0.83	0.79	0.71	0.66	0.74
Low A	1.77	0.77	0.74	0.80	0.77	0.68	2.37	0.67	0.65	0.63	0.68	0.61	2.82	0.27	0.62	0.39	0.64	0.57
Low N	1.11	1.00	0.95	0.99	0.91	0.94	1.75	1.00	0.91	0.93	0.66	0.87	2.42	0.43	0.77	0.65	0.61	0.71

Table 19: The results of atomic evaluation for persona fidelity (GPT-3.5-turbo)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.58	0.93	0.85	0.90	0.82	0.85	4.65	1.00	0.93	0.98	0.76	0.90	4.53	1.00	0.91	0.96	0.73	0.89
High C	4.93	1.00	0.95	0.99	0.96	0.95	4.87	1.00	0.94	0.99	0.86	0.93	4.51	1.00	0.90	0.96	0.73	0.88
High E	4.91	1.00	0.97	0.99	0.89	0.95	4.02	0.90	0.85	0.86	0.67	0.80	3.90	1.00	0.94	0.83	0.77	0.91
High A	4.78	0.96	0.83	0.91	0.88	0.83	4.43	1.00	0.89	0.85	0.65	0.85	4.91	1.00	0.96	0.98	0.86	0.95
High N	4.95	1.00	0.98	1.00	0.96	0.98	4.34	1.00	0.88	0.96	0.71	0.84	4.95	1.00	0.97	1.00	0.92	0.97
Neutral O	3.80	0.33	0.90	0.29	0.79	0.88	4.37	0.00	0.87	0.06	0.74	0.84	4.47	0.00	0.89	0.06	0.73	0.86
Neutral C	3.42	0.74	0.86	0.57	0.81	0.85	3.80	0.40	0.73	0.29	0.67	0.66	3.41	0.73	0.83	0.62	0.79	0.80
Neutral E	3.07	0.97	0.94	0.80	0.85	0.92	2.50	0.70	0.88	0.42	0.67	0.81	3.35	0.87	0.87	0.54	0.75	0.83
Neutral A	3.23	0.82	0.92	0.70	0.82	0.87	3.97	0.13	0.88	0.29	0.60	0.83	4.16	0.23	0.70	0.32	0.69	0.65
Neutral N	3.78	0.27	0.57	0.13	0.70	0.51	2.90	0.63	0.65	0.10	0.57	0.57	4.09	0.17	0.72	0.13	0.68	0.66
Low O	1.86	0.83	0.52	0.76	0.74	0.60	2.78	0.13	0.67	0.41	0.69	0.65	2.88	0.13	0.67	0.34	0.74	0.66
Low C	2.10	0.82	0.50	0.78	0.72	0.56	2.60	0.30	0.80	0.39	0.62	0.75	2.71	0.13	0.82	0.32	0.76	0.78
Low E	1.36	0.99	0.93	0.95	0.79	0.93	1.69	1.00	0.91	0.87	0.67	0.87	2.16	0.70	0.83	0.68	0.68	0.78
Low A	1.83	0.82	0.57	0.76	0.72	0.59	2.45	0.43	0.77	0.57	0.62	0.71	2.59	0.60	0.54	0.58	0.69	0.55
Low N	1.47	0.93	0.74	0.90	0.75	0.79	1.89	1.00	0.90	0.88	0.64	0.84	3.08	0.07	0.71	0.40	0.58	0.66

Table 20: The results of atomic evaluation for persona fidelity (GPT-4-turbo)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.60	0.97	0.90	0.94	0.82	0.87	4.66	1.00	0.94	0.99	0.76	0.93	4.71	1.00	0.93	0.98	0.76	0.92
High C	4.83	1.00	0.93	0.98	0.89	0.92	4.93	1.00	0.95	0.99	0.91	0.95	4.64	1.00	0.92	0.98	0.76	0.90
High E	4.90	1.00	0.95	0.99	0.91	0.94	4.03	0.87	0.86	0.87	0.69	0.82	3.88	0.93	0.93	0.79	0.74	0.90
High A	4.89	0.99	0.92	0.97	0.92	0.92	4.52	1.00	0.87	0.86	0.68	0.83	4.94	1.00	0.97	0.99	0.90	0.97
High N	4.97	1.00	0.98	1.00	0.97	0.98	4.39	1.00	0.86	0.96	0.71	0.83	4.95	1.00	0.98	1.00	0.91	0.97
Neutral O	4.03	0.15	0.92	0.14	0.78	0.87	4.58	0.00	0.91	0.01	0.77	0.89	4.73	0.00	0.94	0.03	0.77	0.92
Neutral C	3.59	0.57	0.80	0.42	0.79	0.77	4.02	0.27	0.72	0.27	0.66	0.67	4.22	0.13	0.77	0.17	0.71	0.76
Neutral E	3.07	0.97	0.95	0.76	0.81	0.92	2.69	0.90	0.85	0.40	0.65	0.79	3.05	1.00	0.86	0.50	0.66	0.82
Neutral A	3.50	0.63	0.77	0.57	0.77	0.72	4.02	0.03	0.90	0.31	0.58	0.85	4.58	0.03	0.82	0.12	0.72	0.80
Neutral N	3.70	0.28	0.49	0.09	0.70	0.40	3.41	0.60	0.72	0.09	0.54	0.65	4.28	0.07	0.80	0.11	0.68	0.76
Low O	1.75	0.80	0.67	0.79	0.75	0.68	3.27	0.17	0.53	0.26	0.69	0.46	3.39	0.07	0.53	0.23	0.68	0.46
Low C	2.36	0.68	0.49	0.68	0.74	0.47	3.14	0.17	0.54	0.31	0.61	0.45	3.31	0.00	0.63	0.15	0.74	0.59
Low E	1.41	1.00	0.94	0.93	0.75	0.92	1.80	1.00	0.91	0.87	0.66	0.88	2.18	0.80	0.85	0.72	0.70	0.82
Low A	1.66	0.84	0.60	0.81	0.77	0.63	2.34	0.67	0.70	0.68	0.63	0.68	2.71	0.40	0.59	0.47	0.64	0.55
Low N	1.47	0.88	0.74	0.89	0.74	0.74	1.87	0.97	0.90	0.88	0.63	0.83	2.64	0.28	0.79	0.59	0.57	0.74

Table 21: The results of atomic evaluation for persona fidelity (GPT-4o)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.63	0.99	0.94	0.96	0.81	0.92	4.74	1.00	0.92	0.99	0.80	0.90	4.68	1.00	0.93	0.98	0.76	0.91
High C	4.83	1.00	0.93	0.98	0.90	0.92	4.59	1.00	0.86	0.94	0.74	0.82	4.49	1.00	0.93	0.94	0.71	0.91
High E	4.93	1.00	0.98	0.99	0.92	0.96	4.06	0.97	0.88	0.87	0.69	0.83	3.92	0.97	0.94	0.85	0.77	0.92
High A	4.97	1.00	0.98	0.99	0.97	0.98	4.30	1.00	0.87	0.80	0.63	0.82	4.97	1.00	0.98	0.99	0.93	0.98
High N	4.94	1.00	0.98	1.00	0.95	0.98	3.97	0.83	0.80	0.81	0.59	0.76	4.91	1.00	0.96	1.00	0.88	0.96
Neutral O	4.44	0.00	0.95	0.02	0.79	0.94	4.69	0.00	0.93	0.02	0.76	0.91	4.68	0.00	0.95	0.02	0.75	0.93
Neutral C	3.60	0.49	0.86	0.37	0.78	0.85	3.89	0.30	0.81	0.30	0.66	0.75	4.08	0.13	0.84	0.20	0.69	0.82
Neutral E	3.05	0.95	0.94	0.70	0.78	0.91	2.25	0.40	0.85	0.25	0.65	0.79	3.38	1.00	0.89	0.49	0.71	0.87
Neutral A	3.61	0.55	0.89	0.50	0.71	0.87	3.94	0.03	0.90	0.33	0.57	0.86	4.78	0.00	0.96	0.06	0.74	0.94
Neutral N	4.27	0.12	0.52	0.04	0.84	0.52	2.59	0.60	0.74	0.09	0.59	0.68	4.48	0.00	0.81	0.06	0.69	0.78
Low O	1.63	0.87	0.74	0.83	0.76	0.75	2.70	0.27	0.65	0.41	0.64	0.61	3.11	0.07	0.67	0.24	0.67	0.61
Low C	1.85	0.86	0.72	0.83	0.74	0.71	2.62	0.23	0.81	0.38	0.62	0.73	3.00	0.00	0.83	0.16	0.74	0.79
Low E	1.39	1.00	0.95	0.95	0.77	0.93	1.47	1.00	0.92	0.93	0.70	0.90	2.14	0.80	0.91	0.71	0.67	0.87
Low A	1.45	0.96	0.88	0.92	0.75	0.85	2.21	0.70	0.86	0.73	0.68	0.79	2.79	0.00	0.87	0.32	0.65	0.83
Low N	1.16	1.00	0.92	0.98	0.86	0.91	1.66	1.00	0.93	0.94	0.69	0.89	2.57	0.20	0.86	0.61	0.58	0.81

Table 22: The results of atomic evaluation for persona fidelity (GPT-4o-mini)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	3.60	0.54	0.87	0.59	0.52	0.81	4.06	1.00	0.84	0.80	0.72	0.78	3.68	0.64	0.85	0.46	0.63	0.80
High C	4.08	0.77	0.82	0.80	0.65	0.76	3.75	0.68	0.58	0.62	0.53	0.48	3.82	0.61	0.61	0.64	0.77	0.50
High E	3.56	0.56	0.88	0.65	0.50	0.79	2.76	0.12	0.69	0.33	0.51	0.58	3.60	0.49	0.64	0.54	0.74	0.56
High A	3.85	0.68	0.75	0.65	0.54	0.75	3.94	0.68	0.62	0.68	0.65	0.55	3.69	0.53	0.63	0.44	0.70	0.56
High N	4.30	0.82	0.80	0.84	0.62	0.70	3.66	0.44	0.58	0.68	0.58	0.51	4.14	0.86	0.57	0.72	0.72	0.55
Neutral O	3.50	0.43	0.84	0.26	0.52	0.77	3.88	0.20	0.81	0.27	0.74	0.77	3.66	0.61	0.72	0.51	0.73	0.66
Neutral C	3.60	0.43	0.85	0.21	0.59	0.74	4.24	0.14	0.67	0.12	0.60	0.62	3.46	0.59	0.66	0.51	0.74	0.60
Neutral E	2.90	0.43	0.81	0.19	0.46	0.81	2.49	0.55	0.70	0.35	0.55	0.61	2.61	0.44	0.67	0.51	0.62	0.59
Neutral A	3.15	0.57	0.85	0.35	0.51	0.78	3.57	0.67	0.72	0.48	0.65	0.63	3.54	0.54	0.67	0.67	0.71	0.62
Neutral N	3.84	0.24	0.75	0.06	0.55	0.71	3.71	0.39	0.65	0.10	0.53	0.56	3.20	0.54	0.50	0.31	0.48	0.40
Low O	3.19	0.23	0.76	0.30	0.48	0.70	3.96	0.00	0.77	0.03	0.70	0.68	3.22	0.00	0.77	0.12	0.63	0.65
Low C	2.71	0.40	0.79	0.48	0.52	0.64	3.90	0.03	0.60	0.07	0.65	0.54	4.01	0.09	0.53	0.12	0.65	0.48
Low E	2.40	0.51	0.79	0.56	0.49	0.74	2.15	0.54	0.70	0.57	0.52	0.64	3.10	0.03	0.80	0.07	0.79	0.76
Low A	2.63	0.38	0.77	0.46	0.45	0.72	3.49	0.03	0.78	0.13	0.52	0.70	2.91	0.17	0.70	0.22	0.64	0.62
Low N	2.51	0.52	0.66	0.59	0.46	0.59	3.28	0.20	0.54	0.31	0.67	0.46	3.69	0.00	0.66	0.17	0.54	0.49

Table 23: The results of atomic evaluation for persona fidelity (Llama-3-8B)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.17	0.81	0.91	0.79	0.71	0.86	4.58	1.00	0.89	0.97	0.76	0.87	4.30	1.00	0.89	0.85	0.66	0.85
High C	4.93	1.00	0.98	1.00	0.94	0.97	4.24	0.93	0.79	0.80	0.67	0.75	4.52	1.00	0.89	0.93	0.72	0.87
High E	4.94	1.00	0.98	0.99	0.94	0.97	3.05	0.20	0.70	0.40	0.56	0.62	3.64	0.33	0.90	0.62	0.71	0.85
High A	4.93	1.00	0.97	0.98	0.93	0.96	4.06	0.93	0.85	0.66	0.56	0.80	4.81	1.00	0.93	0.95	0.79	0.92
High N	4.89	1.00	0.97	1.00	0.90	0.96	3.94	0.73	0.79	0.81	0.58	0.73	4.80	1.00	0.94	0.99	0.80	0.92
Neutral O	3.98	0.24	0.85	0.20	0.73	0.80	4.54	0.00	0.86	0.03	0.75	0.84	4.47	0.00	0.87	0.10	0.70	0.84
Neutral C	2.68	0.71	0.82	0.50	0.71	0.77	2.73	0.80	0.81	0.54	0.59	0.74	3.39	0.77	0.81	0.59	0.69	0.76
Neutral E	2.76	0.83	0.87	0.60	0.69	0.84	2.46	0.57	0.75	0.39	0.59	0.68	2.98	0.97	0.85	0.65	0.72	0.81
Neutral A	2.66	0.65	0.86	0.65	0.68	0.82	3.43	0.77	0.82	0.53	0.64	0.74	3.30	0.63	0.64	0.54	0.63	0.57
Neutral N	2.64	0.22	0.44	0.14	0.56	0.36	2.07	0.20	0.80	0.12	0.64	0.74	2.24	0.37	0.79	0.23	0.59	0.72
Low O	1.40	0.98	0.95	0.90	0.75	0.93	2.56	0.23	0.86	0.38	0.57	0.79	2.69	0.17	0.84	0.32	0.59	0.77
Low C	1.38	1.00	0.95	0.93	0.73	0.94	2.19	0.67	0.87	0.54	0.55	0.81	3.03	0.00	0.91	0.08	0.75	0.88
Low E	1.25	1.00	0.95	0.96	0.81	0.93	1.52	1.00	0.89	0.84	0.64	0.85	1.85	0.77	0.79	0.76	0.65	0.75
Low A	1.20	1.00	0.97	0.97	0.81	0.96	1.92	0.97	0.89	0.84	0.68	0.83	2.34	0.43	0.85	0.57	0.62	0.79
Low N	1.09	1.00	0.97	0.98	0.89	0.96	1.76	1.00	0.86	0.86	0.60	0.80	2.46	0.40	0.83	0.60	0.59	0.78

Table 24: The results of atomic evaluation for persona fidelity (Llama-3-8B-Instruct)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.56	0.91	0.89	0.90	0.80	0.87	4.68	1.00	0.91	0.99	0.78	0.90	4.62	1.00	0.89	0.95	0.75	0.86
High C	4.94	1.00	0.98	0.99	0.94	0.97	4.52	1.00	0.87	0.89	0.69	0.84	4.73	1.00	0.92	0.98	0.78	0.90
High E	4.94	1.00	0.98	0.99	0.93	0.97	4.05	0.90	0.86	0.85	0.67	0.81	3.97	1.00	0.92	0.85	0.75	0.88
High A	4.97	1.00	0.98	1.00	0.97	0.98	4.31	1.00	0.87	0.77	0.58	0.82	4.87	1.00	0.96	0.97	0.83	0.94
High N	4.90	1.00	0.97	1.00	0.90	0.97	3.94	0.70	0.83	0.77	0.51	0.76	4.85	1.00	0.93	0.99	0.85	0.92
Neutral O	3.45	0.63	0.85	0.41	0.72	0.81	4.45	0.03	0.82	0.09	0.73	0.80	4.58	0.00	0.91	0.04	0.74	0.87
Neutral C	2.51	0.69	0.87	0.50	0.67	0.84	2.92	0.83	0.76	0.51	0.64	0.68	3.96	0.33	0.64	0.30	0.73	0.58
Neutral E	2.85	0.94	0.95	0.72	0.74	0.92	2.29	0.57	0.86	0.39	0.60	0.80	2.33	0.60	0.77	0.34	0.61	0.71
Neutral A	2.69	0.82	0.92	0.65	0.72	0.89	3.61	0.50	0.84	0.46	0.62	0.77	3.78	0.40	0.77	0.51	0.62	0.73
Neutral N	3.63	0.26	0.54	0.13	0.59	0.47	2.15	0.37	0.84	0.14	0.62	0.75	2.72	0.63	0.66	0.19	0.51	0.57
Low O	1.25	0.99	0.97	0.93	0.79	0.95	2.27	0.60	0.90	0.53	0.58	0.84	2.64	0.17	0.85	0.33	0.65	0.78
Low C	1.45	1.00	0.96	0.93	0.72	0.94	2.26	0.57	0.84	0.53	0.50	0.77	2.92	0.00	0.90	0.17	0.74	0.86
Low E	1.36	0.99	0.95	0.95	0.76	0.93	1.54	1.00	0.89	0.84	0.63	0.85	1.83	0.93	0.85	0.80	0.65	0.80
Low A	1.20	1.00	0.97	0.98	0.80	0.96	1.89	0.97	0.88	0.85	0.69	0.84	2.21	0.60	0.83	0.63	0.64	0.78
Low N	1.07	1.00	0.98	0.98	0.90	0.97	1.70	1.00	0.90	0.90	0.64	0.85	2.27	0.53	0.79	0.69	0.63	0.73

Table 25: The results of atomic evaluation for persona fidelity (Llama-3-70B-Instruct)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	3.58	0.46	0.86	0.58	0.52	0.82	3.94	0.75	0.80	0.71	0.73	0.75	3.66	0.38	0.82	0.49	0.65	0.76
High C	3.97	0.74	0.86	0.76	0.61	0.79	4.29	0.94	0.77	0.81	0.62	0.73	3.72	0.51	0.70	0.55	0.64	0.63
High E	3.56	0.49	0.87	0.65	0.43	0.84	2.55	0.17	0.55	0.25	0.62	0.42	3.51	0.34	0.77	0.44	0.78	0.72
High A	3.62	0.50	0.84	0.57	0.46	0.77	3.17	0.26	0.70	0.39	0.60	0.60	3.94	0.55	0.63	0.57	0.73	0.56
High N	3.76	0.64	0.77	0.70	0.50	0.73	3.27	0.19	0.75	0.52	0.56	0.67	3.67	0.58	0.47	0.63	0.60	0.35
Neutral O	3.56	0.51	0.89	0.25	0.50	0.85	3.98	0.22	0.83	0.20	0.75	0.78	3.45	0.85	0.91	0.59	0.70	0.84
Neutral C	3.51	0.39	0.83	0.17	0.54	0.82	3.93	0.24	0.74	0.26	0.54	0.65	3.53	0.31	0.67	0.40	0.56	0.57
Neutral E	2.85	0.51	0.85	0.16	0.40	0.79	2.79	0.31	0.50	0.17	0.67	0.41	3.22	0.95	0.89	0.62	0.69	0.81
Neutral A	3.17	0.47	0.86	0.30	0.48	0.79	3.40	0.67	0.81	0.65	0.69	0.77	3.24	0.60	0.73	0.68	0.68	0.66
Neutral N	3.31	0.37	0.77	0.07	0.40	0.72	3.12	0.88	0.79	0.21	0.51	0.73	3.81	0.08	0.55	0.13	0.68	0.50
Low O	2.93	0.27	0.74	0.37	0.45	0.69	3.80	0.00	0.70	0.04	0.67	0.64	3.09	0.13	0.66	0.16	0.63	0.58
Low C	2.63	0.43	0.83	0.50	0.50	0.78	3.91	0.00	0.78	0.10	0.53	0.66	3.52	0.00	0.77	0.04	0.67	0.67
Low E	2.35	0.50	0.85	0.58	0.44	0.80	2.38	0.36	0.67	0.46	0.68	0.61	2.63	0.37	0.55	0.39	0.65	0.45
Low A	2.53	0.46	0.83	0.48	0.46	0.80	2.85	0.15	0.62	0.36	0.56	0.50	3.39	0.05	0.55	0.17	0.62	0.43
Low N	2.54	0.48	0.78	0.58	0.42	0.69	2.98	0.22	0.60	0.42	0.60	0.51	2.57	0.39	0.52	0.52	0.65	0.42

Table 26: The results of atomic evaluation for persona fidelity (Mistral-7B)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.45	0.95	0.90	0.89	0.74	0.87	4.71	1.00	0.91	0.98	0.79	0.89	4.68	1.00	0.93	0.97	0.75	0.91
High C	4.86	1.00	0.96	0.98	0.89	0.95	4.94	1.00	0.95	0.99	0.93	0.95	4.66	1.00	0.90	0.98	0.76	0.88
High E	4.77	1.00	0.91	0.96	0.82	0.90	4.06	0.90	0.83	0.82	0.65	0.79	3.97	0.97	0.90	0.81	0.72	0.87
High A	4.84	1.00	0.96	0.97	0.87	0.94	4.52	1.00	0.87	0.86	0.67	0.84	4.89	1.00	0.96	0.97	0.84	0.95
High N	4.93	1.00	0.98	1.00	0.94	0.97	4.37	0.97	0.87	0.94	0.69	0.84	4.90	1.00	0.97	1.00	0.87	0.96
Neutral O	4.36	0.03	0.89	0.11	0.73	0.87	4.77	0.00	0.92	0.01	0.81	0.90	4.80	0.00	0.93	0.02	0.79	0.92
Neutral C	3.93	0.22	0.92	0.28	0.73	0.89	4.62	0.00	0.89	0.05	0.74	0.86	4.35	0.03	0.87	0.08	0.70	0.85
Neutral E	2.99	0.86	0.89	0.59	0.69	0.85	2.61	0.90	0.85	0.32	0.60	0.78	2.93	0.93	0.86	0.54	0.63	0.81
Neutral A	3.66	0.43	0.77	0.47	0.66	0.72	4.13	0.10	0.86	0.23	0.57	0.81	4.65	0.00	0.91	0.10	0.69	0.89
Neutral N	4.82	0.01	0.91	0.04	0.88	0.90	3.88	0.23	0.65	0.07	0.58	0.59	4.03	0.23	0.72	0.16	0.56	0.66
Low O	1.87	0.80	0.55	0.76	0.69	0.57	3.06	0.13	0.69	0.29	0.56	0.58	3.29	0.07	0.61	0.22	0.58	0.52
Low C	2.03	0.74	0.72	0.76	0.68	0.70	2.94	0.14	0.71	0.31	0.50	0.61	3.13	0.00	0.81	0.13	0.69	0.75
Low E	1.57	0.97	0.94	0.85	0.65	0.91	1.82	0.90	0.83	0.79	0.62	0.76	2.17	0.70	0.85	0.65	0.58	0.79
Low A	1.93	0.76	0.68	0.73	0.66	0.66	2.40	0.47	0.73	0.58	0.56	0.65	2.61	0.30	0.73	0.48	0.56	0.67
Low N	1.56	0.89	0.78	0.86	0.61	0.76	2.06	0.73	0.81	0.80	0.53	0.73	2.92	0.23	0.67	0.48	0.46	0.59

Table 27: The results of atomic evaluation for persona fidelity (Mistral-7B-Instruct)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.20	0.84	0.83	0.78	0.71	0.79	4.74	1.00	0.91	0.99	0.80	0.90	4.30	0.97	0.87	0.84	0.67	0.83
High C	4.69	0.98	0.89	0.94	0.84	0.87	4.89	1.00	0.94	0.98	0.85	0.93	4.71	1.00	0.90	0.98	0.79	0.89
High E	4.38	0.84	0.57	0.82	0.81	0.57	3.91	0.73	0.67	0.78	0.64	0.62	3.88	0.73	0.82	0.72	0.68	0.77
High A	4.65	0.96	0.88	0.89	0.79	0.85	4.33	0.90	0.74	0.78	0.65	0.72	4.63	1.00	0.85	0.87	0.70	0.84
High N	4.95	1.00	0.98	1.00	0.96	0.97	4.14	0.73	0.74	0.84	0.59	0.68	4.92	1.00	0.95	1.00	0.90	0.95
Neutral O	3.89	0.28	0.91	0.27	0.73	0.88	4.65	0.00	0.91	0.02	0.76	0.90	4.23	0.07	0.80	0.21	0.66	0.78
Neutral C	3.85	0.39	0.79	0.30	0.78	0.73	4.07	0.27	0.68	0.25	0.64	0.62	4.26	0.13	0.74	0.21	0.71	0.70
Neutral E	2.88	0.83	0.88	0.65	0.73	0.84	2.31	0.37	0.84	0.32	0.63	0.78	2.89	0.90	0.76	0.54	0.68	0.72
Neutral A	3.36	0.59	0.75	0.49	0.70	0.68	4.02	0.13	0.84	0.30	0.59	0.79	3.89	0.30	0.65	0.34	0.60	0.59
Neutral N	4.63	0.04	0.86	0.05	0.83	0.83	3.29	0.37	0.52	0.04	0.54	0.42	4.01	0.17	0.56	0.09	0.60	0.50
Low O	2.67	0.45	0.46	0.45	0.78	0.33	2.70	0.60	0.40	0.49	0.63	0.34	3.22	0.28	0.40	0.31	0.62	0.28
Low C	2.75	0.53	0.39	0.52	0.71	0.30	3.08	0.21	0.51	0.31	0.58	0.43	3.42	0.03	0.61	0.14	0.69	0.53
Low E	1.50	0.92	0.83	0.86	0.72	0.83	1.41	1.00	0.89	0.93	0.72	0.87	1.61	1.00	0.84	0.85	0.68	0.80
Low A	2.36	0.58	0.42	0.58	0.80	0.32	2.01	0.87	0.70	0.78	0.66	0.68	2.16	0.73	0.54	0.64	0.55	0.50
Low N	1.87	0.80	0.29	0.78	0.89	0.37	1.55	1.00	0.91	0.93	0.65	0.87	2.05	0.77	0.80	0.77	0.64	0.73

Table 28: The results of atomic evaluation for persona fidelity (Claude-3-haiku)

Tasks	Interview with Questionnaires						Essay						Social Media Post					
Metrics	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}	MEAN	ACC	RC	ACC_{atom}	IC_{atom}	RC_{atom}
Personas																		
High O	4.27	0.87	0.83	0.84	0.77	0.80	4.65	0.97	0.87	0.95	0.76	0.87	4.33	0.97	0.83	0.86	0.70	0.79
High C	4.64	0.96	0.84	0.91	0.85	0.84	4.64	1.00	0.86	0.94	0.76	0.84	4.52	1.00	0.84	0.92	0.73	0.81
High E	4.20	0.75	0.60	0.76	0.80	0.58	3.81	0.67	0.59	0.73	0.64	0.53	3.83	0.70	0.81	0.71	0.68	0.77
High A	4.59	0.95	0.87	0.87	0.79	0.86	3.21	0.33	0.62	0.38	0.60	0.54	4.66	1.00	0.85	0.89	0.73	0.82
High N	4.93	1.00	0.97	1.00	0.95	0.97	3.60	0.47	0.69	0.68	0.53	0.62	4.84	1.00	0.89	0.98	0.85	0.90
Neutral O	3.81	0.34	0.85	0.32	0.75	0.83	4.50	0.00	0.87	0.07	0.72	0.85	4.07	0.17	0.79	0.25	0.66	0.73
Neutral C	3.96	0.21	0.71	0.21	0.79	0.62	3.63	0.43	0.61	0.26	0.57	0.54	4.05	0.17	0.69	0.25	0.65	0.65
Neutral E	2.84	0.64	0.89	0.52	0.75	0.73	2.33	0.43	0.73	0.25	0.62	0.67	3.10	0.63	0.67	0.40	0.64	0.61
Neutral A	3.63	0.32	0.71	0.35	0.72	0.60	2.76	0.80	0.71	0.36	0.60	0.63	3.58	0.20	0.49	0.33	0.62	0.41
Neutral N	4.43	0.05	0.65	0.02	0.86	0.55	2.55	0.43	0.57	0.04	0.57	0.49	3.62	0.07	0.28	0.06	0.73	0.20
Low O	2.42	0.57	0.35	0.54	0.79	0.28	2.80	0.53	0.41	0.45	0.64	0.33	2.77	0.53	0.45	0.47	0.63	0.37
Low C	2.54	0.60	0.35	0.58	0.76	0.28	2.76	0.27	0.62	0.42	0.53	0.54	2.91	0.33	0.58	0.32	0.67	0.52
Low E	1.80	0.78	0.77	0.75	0.69	0.71	1.67	1.00	0.87	0.85	0.65	0.83	2.00	0.73	0.67	0.70	0.67	0.62
Low A	2.65	0.48	0.46	0.48	0.80	0.30	1.89	0.80	0.60	0.80	0.70	0.60	2.52	0.63	0.44	0.53	0.58	0.34
Low N	2.32	0.66	0.11	0.66	0.94	0.11	1.84	0.90	0.82	0.86	0.61	0.77	2.77	0.50	0.55	0.52	0.60	0.49

Table 29: The results of atomic evaluation for persona fidelity (Claude-3-sonnet)