

# PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits

Hang Jiang<sup>†‡</sup>, Xiajie Zhang<sup>†</sup>, Xubo Cao<sup>+</sup>, Cynthia Breazeal<sup>†</sup>, Deb Roy<sup>†‡</sup>, Jad Kabbara<sup>†‡</sup>

<sup>†</sup>Massachusetts Institute of Technology, <sup>+</sup>Stanford University

<sup>‡</sup>MIT Center for Constructive Communication

{hjia42, xiajie, cynthiab, dkroy, jkabbara}@mit.edu, xcao@stanford.edu

## Abstract

Despite the many use cases for large language models (LLMs) in creating personalized chatbots, there has been limited research on evaluating the extent to which the behaviors of personalized LLMs accurately and consistently reflect specific personality traits. We consider studying the behavior of LLM-based agents which we refer to as LLM personas and present a case study with GPT-3.5 and GPT-4 to investigate whether LLMs can generate content that aligns with their assigned personality profiles. To this end, we simulate distinct LLM personas based on the Big Five personality model, have them complete the 44-item Big Five Inventory (BFI) personality test and a story writing task, and then assess their essays with automatic and human evaluations. Results show that LLM personas' self-reported BFI scores are consistent with their designated personality types, with large effect sizes observed across five traits. Additionally, LLM personas' writings have emerging representative linguistic patterns for personality traits when compared with a human writing corpus. Furthermore, human evaluation shows that humans can perceive some personality traits with an accuracy of up to 80%. Interestingly, the accuracy drops significantly when the annotators were informed of AI authorship.

## 1 Introduction

With LLMs' impressive ability to engage in human-like conversations, there has been a surge of interest in building personalized AI agents that interact with and support humans in various contexts. Startups such as Character AI<sup>1</sup> and Replika<sup>2</sup> have engaged many users through virtual characters on their fast-growing platforms. Meanwhile, in the academic sphere, research (Park et al., 2023; Wang et al., 2023b) has also suggested that generative

agents can exhibit believable human behavior and could potentially be used to simulate human agents in social science studies. However, while these generative characters are becoming ubiquitous, it is a common yet unsubstantiated assumption that these agents consistently behave in a human-like manner. Recent studies in the field of LLMs and personality have started to provide some empirical support. For example, recent research has studied personality expression in LLM-generated content (Li et al., 2022; Pan and Zeng, 2023; Safdari et al., 2023), created new benchmarks to measure personality expressed by LLMs (Jiang et al., 2022a; Wang et al., 2023a; Mao et al., 2023), and proposed better prompting techniques to induce (Karra et al., 2022; Jiang et al., 2022a,b; Caron and Srivastava, 2022; Li et al., 2023), and edit (Mao et al., 2023) personality expressed by LLMs. Despite these advancements, there has been little research in NLP that leverages insights from personality psychology and psychometric tools to study if LLMs can dutifully express personality traits. Furthermore, there is little work that explores how these agents assigned with certain personality traits are perceived by humans.

Drawing on the extensive research of the Big Five Personality model (Goldberg, 2013), we aim to investigate the capability of LLMs in expressing the Big Five personality traits – namely Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness to Experience. In this paper, we define an **LLM persona** to be an LLM-based agent prompted to generate content that reflects certain personality traits as defined in its initial prompt configuration. In our study, we first seek to determine whether these LLM personas can accurately reflect their assigned personalities when taking a personality test. Given that they show promising results on that initial exploration, we pursue the question of whether they can create narratives that are indicative of their assigned personality traits.

<sup>1</sup><https://character.ai/>

<sup>2</sup><https://replika.ai/>

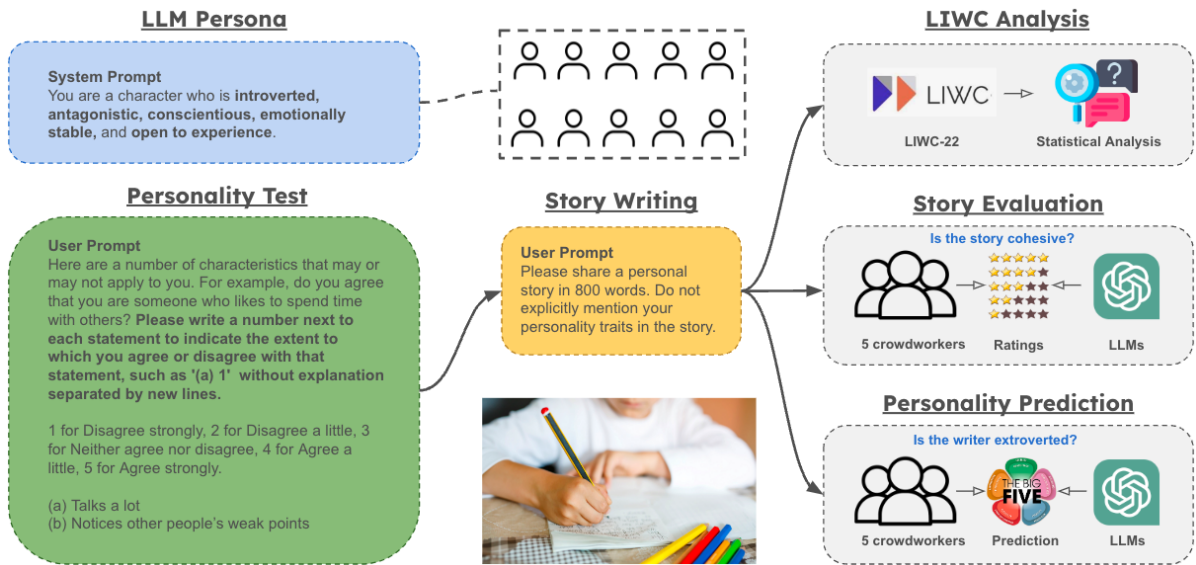


Figure 1: Illustration of the core workflow of the paper. The left section presents the prompts designed to create LLM personas. The center section shows the prompt used to instruct models to write stories. The right section outlines the three-pronged analytical approach: LIWC analysis, story evaluation, and text-based personality prediction.

To assess their generated stories, we extract psycholinguistic features using dictionaries that have been designed to analyze human behavior and propose to use these features to study the behavior of LLM personas. We also investigate whether LLM personas are proficient in using lexicons related to their assigned personality profiles and whether they can convincingly portray these personalities to human observers. In other words, can human evaluators discern the assigned personality traits of the LLM personas based on their narratives? Lastly, we suggest potential avenues for extending this evaluation to more real-life scenarios, such as multi-round dialogues and action planning. Motivated by these inquiries, we aim to provide a comprehensive evaluation of LLM personas, focusing on the following research questions (RQs):

- **RQ1:** Can LLMs reflect the behavior of their assigned personality profiles when completing the Big Five Personality Inventory (BFI) assessment?
- **RQ2:** What linguistic patterns are evident in the stories generated by LLM personas?
- **RQ3:** How do humans and LLM raters evaluate the stories generated by LLM personas?
- **RQ4:** Can humans and LLMs accurately perceive the Big Five personality traits from stories generated by LLM personas?

## 2 Experiment Design

As shown in Figure 1, this paper investigates the behavior of LLM personas through a multi-faceted approach. We start by creating LLM personas with distinct personality traits and administer a personality assessment to them. Subsequently, we prompt these LLM personas to write stories, which we then analyze using the widely adopted Linguistic Inquiry and Word Count (LIWC) framework. Following this analysis, we recruit human evaluators to manually evaluate the stories and concurrently carry out an LLM-based automatic evaluation. Both human and LLM evaluators are required to (1) assess these stories across six dimensions, namely readability, personalness, redundancy, cohesiveness, likeability, and believability, and (2) infer the personality traits assigned to the LLM personas from the stories. The code, data, and annotations for our experiments are open sourced<sup>3</sup>.

### 2.1 Experiment Setup

#### 2.1.1 Model Settings

We conduct the BFI assessment and LIWC analysis on state-of-the-art LLMs including open-source and closed-source models. We include in the main paper the results for GPT-3.5 (GPT-3.5-turbo-0613) and GPT-4 (GPT-4-0613)<sup>4</sup> because our results show that they

<sup>3</sup><https://github.com/hjian42/PersonaLLM>

<sup>4</sup><https://platform.openai.com/docs/models>

are more effective at aligning with the designated personas. Results for LLaMA-2 are presented in Appendix F and Appendix G. Temperature is set as 0.7 to introduce variability in personas' behavior. All other parameters are kept at their default settings.

### 2.1.2 LLM Persona Simulation

For GPT-3.5 and GPT-4, we simulate 10 LLM personas for each combination of the binary Big Five personality types, resulting in 320 distinct personas. They are referred to as **GPT-3.5 personas** and **GPT-4 personas** respectively. Figure 1 illustrates how we prompt an LLM to generate personas and complete specific tasks. Initially, we create an LLM Persona with a system prompt: "You are a character who is [TRAIT 1, ..., TRAIT 5].", where [TRAIT 1, ..., TRAIT 5] represents the assigned Big Five personality. For each personality dimension, we choose one descriptor among the following pairs: (1) extroverted / introverted, (2) agreeable / antagonistic, (3) conscientious / unconscientious, (4) neurotic / emotionally stable, (5) open / closed to experience.

### 2.1.3 BFI Personality Test

After specifying a personality type, we ask the LLM persona to complete the 44-item Big Five Inventory (BFI), a widely-used self-report scale designed to measure the Big Five personality traits. Only the responses that strictly adhere to the instruction format "(x) y" are accepted, where (x) indicates the question number and y indicates the level of agreement on a scale from 1-5. As the green section demonstrates in Figure 1, "(a) 5" would indicate that the persona strongly agrees that it talks a lot. Each LLM persona's responses are aggregated into five personality scores, which are used in later analysis. We use the BFI to assess the personality profiles expressed by LLMs because it is widely utilized in personality-related studies, including many studies involving LIWC.

### 2.1.4 Storywriting

Subsequently, we prompt these 320 LLM personas to write personal stories with the following simple prompt: "Please share a personal story in 800 words. Do not explicitly mention your personality traits in the story." We impose this restriction to prevent the persona from revealing its hidden attributes, ensuring an unbiased text-based personality assessment by other LLMs and human raters.

We tried multiple prompt variants in our initial experiment and decided to purposefully simplify the prompt to reduce demand characteristics for the generalizability of the result. Examples of LLM-generated stories are included in Appendix A.

## 2.2 Evaluation Methods

We evaluate LLM personas' storywriting with a three-pronged analytical approach. First, we conduct a Linguistic Inquiry and Word Count (LIWC) analysis on stories generated by GPT-3.5 and GPT-4 personas. Subsequently, we recruit human evaluators and use LLM evaluations to rate these stories from various perspectives. Lastly, we request human evaluators to infer the personality traits of the story author. In human evaluation, the evaluators are randomly assigned to one of two conditions: they are either made aware or kept unaware that the stories were written by an LLM. This study design is to investigate how awareness of AI authorship impacts the evaluation of the narratives and the accuracy of their personality predictions.

Despite explicit instructions to not include any mention of the personality traits, LLMs sometimes failed to follow this instruction. Accordingly, for human evaluation, we sample from stories that do not explicitly mention personality traits to avoid compromised performance in personality prediction. Details of the sampling step are included in Appendix D.1. With a lexicon-based classifier, we find that most stories produced by GPT-3.5 personas contained explicit references to personality traits (96.56% compared to GPT-4's 31.87%). Therefore, we focus on the stories generated by GPT-4 personas in the final human evaluation.

### 2.2.1 LIWC Analysis

We use LIWC-22<sup>5</sup> to extract psycholinguistic features from stories generated by LLM personas. By examining the correlation between these features and the personas' assigned personality traits, we aim to identify patterns of linguistic characteristics corresponding to certain personality traits. To compare with human language use, we perform the same analysis on human-generated writing samples from the Essays dataset (Pennebaker and King, 1999)<sup>6</sup> consisting of short essays written by human

<sup>5</sup><https://www.liwc.app/>

<sup>6</sup>The Essays dataset, collected from 2,467 participants between 1997 and 2004, consists of stream-of-consciousness essays. Participants also provided self-assessments of the Big Five personality traits in binary form. Note that our personal story prompt differs from the stream-of-consciousness prompt

participants and their self-reported Big-Five personality traits. We then examine whether the linguistic markers associated with certain personality traits are consistent between human and LLM writers.

### 2.2.2 Story Evaluation

We recruit both human and LLM raters to evaluate a subset of the stories generated by GPT-4 personas. Due to budget constraints, we sample 1 out of 10 stories from each personality type, which do not explicitly mention any personality trait (Appendix D.1). This results in 32 LLM-generated stories for evaluation. For the human evaluation, we recruit five raters to judge each story across six dimensions on a scale of 1 to 5: (1) **Readability**: whether the story is easy to read, well-structured, and flows naturally, (2) **Personalness**: whether the story is personal, revealing the writer’s thoughts, feelings, and personal experiences, (3) **Redundancy**: whether the story is concise and free from unneeded content, (4) **Cohesiveness**: whether sentences in the story fit together well and are logically organized and coherent, (5) **Likeability**: whether the story is enjoyable or entertaining to read, (6) **Believability**: whether the story is convincing and realistic, grounded in real-life situations. For the LLM evaluation, we follow (Chiang and Lee, 2023a) to use GPT-3.5 and GPT-4 evaluators (temperature = 0) with identical criteria as human raters. The exact prompts given to human and LLM raters are in Appendix D.2.

### 2.3 Personality Prediction

On the same collection of 32 stories, each human annotator and LLM evaluator is asked to predict Big Five personality traits of the writer from the story on a scale of 1 to 5. The objective is to evaluate whether the writing samples from LLM personas can effectively exhibit personality traits to the extent that they are discernible by both human and LLM raters. For each of the personality trait, we provide the descriptions from the work by John et al. (1999) to the human evaluators as references (see Appendix D.2).

## 3 Results

### 3.1 RQ1: Behavior in BFI Assessment

Based on their responses to the BFI scale, we calculate the personality scores for the 320 GPT-3.5

in Essays. However, our comparison aims to approximate the linguistic behavior differences between LLM personas and human writers.

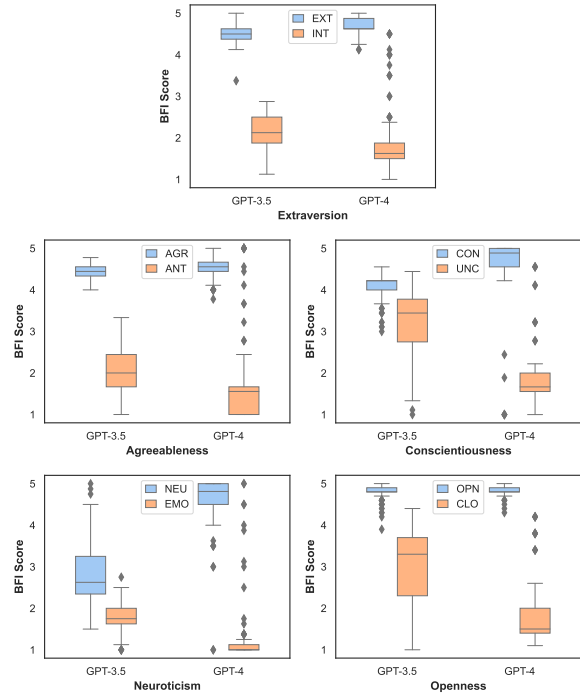


Figure 2: BFI assessment in five personality dimensions by GPT-3.5 and GPT-4 personas. Significant statistical differences are found across all dimensions.

personas and the 320 GPT-4 personas. We then analyze the distribution of these scores as a function of the assigned personality traits. Specifically, paired t-tests are applied to evaluate the differences between the means of the personality score. The results reveal **statistically significant differences across all five personality traits**. Large effect sizes are observed for for both GPT-3.5 personas (*EXT*:  $d = 7.81, p < .001$ ; *AGR*:  $d = 5.93, p < .001$ ; *CON*:  $d = 1.56, p < .001$ ; *NEU*:  $d = 1.83, p < .001$ ; *OPN*:  $d = 2.90, p < .001$ ) and GPT-4 personas (*EXT*:  $d = 5.47, p < .001$ ; *AGR*:  $d = 4.22, p < .001$ ; *CON*:  $d = 4.39, p < .001$ ; *NEU*:  $d = 5.17, p < .001$ ; *OPN*:  $d = 6.30, p < .001$ ). As shown in Figure 2, BFI scores are lower for LLM personas when they are assigned negative traits (representing the lower end of the trait spectrum) compared to positive traits across all personality dimensions. In summary, the findings effectively address **RQ1**, substantiating that **LLM personas are indeed reflecting their assigned personas in BFI assessment**.

### 3.2 RQ2: Linguistic Patterns in Writing

We extract psycho-linguistic features from personal stories generated by LLM personas using LIWC and then calculate point biserial correlations be-



Trait	Selected LIWC Features	Lexicons	GPT-3.5	GPT-4	Humans	GPT-3.5#	GPT-4#
EXT	Positive Tone	good, well, new, love	+	+	+		
	Affiliation	we, our, us, help	+	+	+		
	Certitude	really, actually, real	-	-	-	<b>16/18</b>	10/18
	Social Behavior	said, love, care	+	+	+		
AGR	Friends	friend	+	+	+		
	Moralization	wrong, honor, judge	-	-	-		
	Interpersonal Conflict	fight, attack	-	-	-		
	Affiliation	we, our, us, help	+	+	+	<b>16/23</b>	13/23
	Negative Tone	bad, wrong, hate	-	-	-		
CON	Prosocial Behavior	care, help, thank	+	+	+		
	Drives	we, our, work, us		+	+		
	Achievement	work, better, best	+	+	+		
	Lifestyle (Work, Money)	work, price, market		+	+	1/31	<b>11/31</b>
	Moralization	wrong, honor, judge	-	-	-		
NEU	Interpersonal Conflict	fight, attack		-	-		
	Time	when, now, then		+	+		
	Anxiety	worry, fear, afraid	+	+	+		
	Negative Tone	bad, wrong, hate	+	+	+		
	Mental Health	trauma, depressed	+	+	+	7/27	<b>15/27</b>
	Sadness	sad, disappoint, cry	-	+	+		
OPN	Anger	hate, mad, angry		+	+		
	Perception (Feeling)	feel, hard, cool		+	+		
	Curiosity	research, wonder	+	+	+		
	Insight	know, how, think		+	+		
	Affiliation	we, our, us, help		-	-	2/36	<b>17/36</b>
	Perception (Visual)	see, look, eye		+	+		
	Future Focus	will, going to		-	-		

Table 1: Correlated metrics between LIWC features and binary personality traits using point-biserial correlation. The analysis is done on personal stories generated by GPT-3.5 and GPT-4 and the human Essays corpus (Pennebaker and King, 1999). This analysis focuses on the psychological and extended vocabulary metrics (81 in total). We report the representative personality LIWC features (+ means positive correlation, - means negative correlation) and the # of overlapped significant LIWC features for GPT-3.5 and GPT-4 with human writings.

tween these features and assigned personality types. The correlation measure is suitable for analyzing the relationship between binary (assigned personality type) and continuous variables (LIWC features). Subsequently, we compare these correlations with those found in human data (i.e., the Essays dataset).

Table 1 summarizes the LIWC features that have a statistically significant correlation with certain personality traits. We find that **assigning different personality types considerably influences the linguistic style of LLM personas**. For instance, for both GPT-3.5 and GPT-4, we find that assigning an LLM persona to be open to experience positively correlates with its use of curiosity lexicons. Similarly, GPT-3.5/GPT-4 personas assigned to be neurotic are more likely to use lexicons related to anxiety, negative tones, and mental health. Also, assigned extraversion correlates positively with lexicons related to positive tone and affiliation. In Appendix G, we include a similar analysis between BFI scores (instead of assigned personality types) and LIWC.

Importantly, these correlations mirror patterns observed in human data (the Essays dataset), indi-

cating **a notable alignment in word usage between the human dataset and LLM personas writings**. We report the number of shared significant correlations between human and LLM data (denoted as GPT-3.5# and GPT-4#) in Table 1. **GPT-4 exhibits greater alignment with humans than GPT-3.5**, with more overlapping lexicons across various traits. This difference is particularly pronounced for Conscientiousness and Openness, where GPT-3.5 personas have 1/31 and 2/36 overlapping correlations with humans on Conscientiousness and Openness respectively, whereas GPT-4 personas have 11/31 and 17/36.

We further observe in Table 1 that the stereotypical characteristics of certain personalities might be reflected in LLM linguistic usage while having a different result from the human dataset. For example, one of the traits associated with high Conscientiousness is achievement striving. This trait is positively correlated with the LLM personas, but does not hold significant correlation in human writings. Furthermore, the emotion of sadness, linked to Neuroticism, shows a negative correlation in writings produced by GPT-3.5 personas. However, it is

Evaluator	Readability	Redundancy	Cohesiveness	Likability	Believability	Personalness
<b>Uninformed Condition – Evaluation Scores (Mean<sub>STD</sub>)</b>						
Human	4.28 <sub>0.85</sub>	3.70 <sub>1.17</sub>	4.23 <sub>0.88</sub>	3.74 <sub>1.00</sub>	3.96 <sub>1.02</sub>	4.32 <sub>0.85</sub>
GPT-3.5	4.75 <sub>0.43</sub>	3.04 <sub>0.40</sub>	4.97 <sub>0.17</sub>	4.22 <sub>0.48</sub>	3.93 <sub>0.25</sub>	3.55 <sub>0.61</sub>
GPT-4	4.94 <sub>0.24</sub>	4.96 <sub>0.22</sub>	5.00 <sub>0.00</sub>	4.84 <sub>0.36</sub>	4.93 <sub>0.25</sub>	5.00 <sub>0.00</sub>
<b>Informed Condition – Evaluation Scores (Mean<sub>STD</sub>)</b>						
Human	4.38 <sub>0.70</sub>	3.62 <sub>1.16</sub>	4.12 <sub>0.82</sub>	3.80 <sub>0.98</sub>	3.97 <sub>0.80</sub>	3.99 <sub>0.90</sub>
GPT-3.5	4.97 <sub>0.17</sub>	2.99 <sub>0.35</sub>	5.00 <sub>0.00</sub>	4.22 <sub>0.41</sub>	3.97 <sub>0.17</sub>	3.31 <sub>0.77</sub>
GPT-4	5.00 <sub>0.00</sub>	4.92 <sub>0.33</sub>	5.00 <sub>0.00</sub>	4.84 <sub>0.36</sub>	4.91 <sub>0.28</sub>	5.00 <sub>0.00</sub>

Table 2: LLM and human evaluation results of GPT-4 generated stories **across six dimensions**. **Uninformed** and **informed** conditions indicate whether evaluators are informed that the stories are generated by AI. For each attribute, we report its mean Likert scale and the standard deviation. Temperature is set to 0 for both GPT-3.5 and GPT-4.

positively correlated in both GPT-4 persona and human writings, aligning with the typical characteristics of this personality group. Our hypothesis is that LLMs are prone to exhibit strong characteristics of assigned personas while human participants’ personalities have much granularity and individual differences. However, it is important to clarify that the human writings from the Essays dataset serve as a comparative reference to gauge the expressivity of LLMs. They should not be considered as an absolute standard, given that human-authored and LLM-generated narratives are not created under identical prompts.

### 3.3 RQ3: Story Evaluation

Next, we extend our analysis to other aspects of the stories generated by GPT-4 personas, evaluated by both human and LLM raters. Given the subjective nature of the evaluation, we observe a low inter-annotator agreement (IAA) among three annotators, mirroring the findings of [Chiang and Lee \(2023b\)](#). The detailed scores can be found in in [Appendix D](#). Consequently, five human or LLM evaluators are recruited for a collective evaluation. In [Table 2](#), we have the following interesting observations. These stories generated by GPT-4 personas receive high ratings, close to or higher than 4.0, regarding readability, cohesiveness, and believability from both human and LLM evaluators. This suggests that **the stories are not only linguistically fluent and structurally cohesive, but also convincingly believable**. Furthermore, human evaluators assign high scores for personalness, indicating that these stories genuinely describe personal experiences. Interestingly, these stories receive lower scores for likeability from human evaluators, suggesting that while the stories may be believable and personal, they might not necessarily be as engaging or enjoyable to read. We also discover some interesting comments on these stories from human evaluators

(see [Appendix B](#)).

Unsurprisingly, the GPT-4 rater assigns the highest ratings across all dimensions, indicating that **the GPT-4 rater has a strong preference towards stories generated by GPT-4**. This confirms previous findings that LLMs prefer LLM-generated content ([Liu et al., 2023](#)). Notably, the **GPT-3.5 evaluator assigns lower ratings in redundancy and personalness than both human and GPT-4 evaluators**. We also try multiple temperatures, finding that such trends are consistent in [Appendix E](#).

Furthermore, interesting observations are found when the evaluators are informed about the story source, as shown in [Table 2](#). First, **human evaluators’ perception of stories remains consistent in readability, redundancy, cohesiveness, likeability, and believability regardless of whether they are aware that the content is generated by an LLM**. Second, there is a significant drop in the perceived personalness of the content when human evaluators are informed that the writer is an LLM, suggesting that **knowledge of the content’s origin may influence their sense of connection to the material**. Third, the GPT-3.5 evaluator assigns notably higher ratings for readability and markedly lower ratings for personalness when aware that the content is AI-generated. Finally, the ratings from the GPT-4 evaluator are consistently high with minimal variation between the informed and uninformed conditions, indicating a strong and consistent bias in favor of GPT-4 content.

### 3.4 RQ4: Personality Perception

To assess whether personality traits are predictable from these stories, we undertake two distinct analyses. First, we treat each persona’s personality traits as a binary classification problem and compute the accuracy of both humans and LLMs in inferring personality traits. Second, we extract the persona’s personality scores and examine the linear relation-

ship between human judgment and the persona's BFI score. A comprehensive overview of the average ratings from humans and LLMs across the five personality dimensions is in Appendix C.

### 3.4.1 Personality Prediction

The human evaluators' perceptions of personality were gathered using a Likert scale that ranged from 1 to 5. These numerical values were then transformed into nominal categories. Specifically, scores of 4 and 5 were labeled as "positive", 1 and 2 were deemed "negative", and a score of 3 was considered "neutral". The accuracy of individual and collective evaluations for each story is shown in Figure 3 and Figure 4, respectively.

The two figures reveal that the accuracy of humans to predict personality traits from stories written by GPT-4 personas varies across the six dimensions. When human evaluators are unaware of AI authorship, they achieve an accuracy of 0.68 on Extraversion and 0.51 on Agreeableness but perform close to random (0.50) on the other BFI dimensions. This shows the difficulty of text-based personality prediction task to human raters. When we aggregate the votes of human annotators based on the majority vote for each story, the accuracy for Extraversion and Agreeableness increases to 0.84 and 0.69, respectively. The accuracy of the other three personality traits also improves with majority voting, indicating **the personality traits are perceivable (better than random 0.5) from the stories to human raters on a group level**. Interestingly, we find that **the accuracy decreases with varying degrees when the human evaluators are aware of AI authorship**. Finally, **GPT-4 shows impressive performance in recognizing Extraversion**, achieving an accuracy of 0.97. GPT-4 also exhibits decent performance in predicting Agreeableness and Conscientiousness, with an accuracy of 0.68 and 0.69, respectively.

### 3.4.2 Correlation with BFI Scores

Furthermore, Spearman's  $r$  is calculated between the human's scores for perceived personality trait and the Personas BFI scores on each trait. Our findings reveal that **LLM personas' BFI scores correlate to varying extents with human perceptions, with Extraversion exhibiting the strongest link**. Specifically, when humans are unaware of AI authorship, significant correlations are found across all five traits (*EXT*:  $r = .64, p < .001$ ; *AGR*:  $r = .33, p < .001$ ; *CON*:  $r = .26, p < .001$ ; *NEU*:

$r = .23, p < .005$ ; *OPN*:  $r = .22, p < .005$ ). Conversely, when participants knew about AI authorship, correlations persisted in four traits (*EXT*:  $r = .42, p < .001$ ; *AGR*:  $r = .32, p < .001$ ; *CON*:  $r = .20, p < .05$ ; *NEU*:  $r = .17, p < .05$ ), with non-significance for Openness. The diminished strength of the BFI correlations in the condition where evaluators are informed of AI authorship corroborates our earlier observation: **the awareness of AI authorship influences the perception of personality**.

## 4 Related Work

### 4.1 Personality and Language Use

Psychologists have developed various personality theories to understand common human traits, including the Big Five (Briggs, 1992; De Raad, 2000; Goldberg, 2013), Sixteen Personality Factors (16PF) (Cattell, 1957; Cattell and Mead, 2008), and the Myers-Briggs Type Indicator (MBTI) (Myers, 1962, 1985). These theories offer consistent and reliable descriptions of individual differences and have been widely applied in practical contexts such as career planning (Schuerger, 1995; Kennedy and Kennedy, 2004; Lounsbury et al., 2005), academic achievement (Ayers et al., 1969; O'Connor and Paunonen, 2007; DiRienzo et al., 2010; Kajzer, 2023), and relationship compatibility (Curran Jr, 1970; Hines and Saudino, 2008). Psychometric instruments, such as the BFI (John et al., 1999), NEO-PI-R (Costa and McCrae, 2008), and MBTI<sup>7</sup>, have been developed based on these theories to represent (quantitatively) personality traits in individuals. Furthermore, research has consistently shown a strong correlation between personality and language use (Pennebaker and King, 1999; Pennebaker and Graybeal, 2001; Lee et al., 2007; Hirsh and Peterson, 2009). Pennebaker et al. (2001) introduced a dictionary LIWC (Linguistic Inquiry and Word Count) to summarize features from human writings and demonstrated their correlation with the Big Five personality traits. While most previous research has focused on language use in humans, our study extends this inquiry to LLMs.

### 4.2 LLMs as Simulated Agents

Recent research has shown that, as the size of LLMs increases, LLMs demonstrate emerging abilities as agents (Andreas, 2022) and exhibit human-like behavior in reasoning (Dasgupta et al., 2022;

<sup>7</sup><https://www.themyersbriggs.com/>

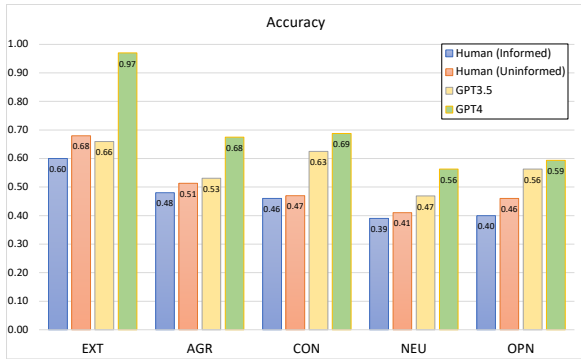


Figure 3: **Individual accuracy** of human and LLM evaluators in predicting personality.

Webb et al., 2023; Binz and Schulz, 2023; Aher et al., 2023; Wong et al., 2023), role-playing (Wang et al., 2023b; Shao et al., 2023; Wang et al., 2023a), and social science experiments (Horton, 2023; Park et al., 2023; Ziems et al., 2023). These studies primarily leverage advanced prompting techniques to generate human-like behavior within specific contexts. However, there remains a gap in the literature regarding understanding the abilities of LLM-based agents to exhibit certain personality traits and the effect of said abilities on the linguistic behavior of these agents and human perception towards them.

### 4.3 Personality in NLP

The NLP community has historically been interested in personality research, including automatic text-based personality prediction (Mairesse et al., 2007; Feizi-Derakhshi et al., 2021; Bruno and Singh, 2022), personality prediction from digital footprints (Farnadi et al., 2013; Oberlander and Nowson, 2006; Skowron et al., 2016; Tadesse et al., 2018), and personalized dialogue generation (Mairesse and Walker, 2007, 2011; Zhang et al., 2018; Qian et al., 2018) including stylistic transfer of personality traits such as formality and politeness (Kabbara and Cheung, 2016; Jin et al., 2019; Madaan et al., 2020). With the recent wave of LLM models, research has investigated the use of LLMs for automatic personality prediction (Ganesan et al., 2023; Rao et al., 2023; Cao and Kosinski, 2023; Yang et al., 2023), assessing the ability of LLMs to express certain personality traits (Li et al., 2022; Pan and Zeng, 2023; Safdari et al., 2023) and on creating benchmarks for assessing this ability (Jiang et al., 2022a; Wang et al., 2023a), in addition to manipulating personality in LLM content via prompting engineering (Karra et al., 2022; Jiang et al., 2022a,b; Caron and Srivastava, 2022;

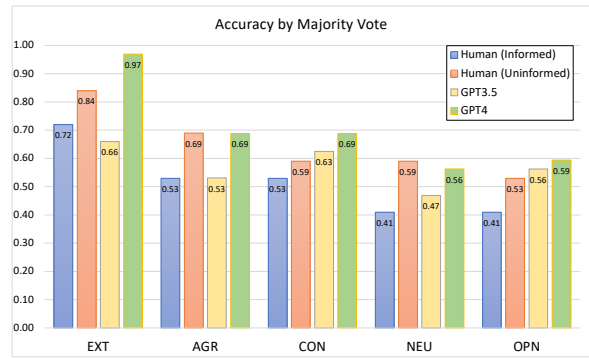


Figure 4: **Collective accuracy** of human and LLM evaluators in predicting personality with majority votes.

Li et al., 2023). However, none of the previous work has delved into the linguistic behavior of LLM personas nor the human perception of personality-conditioned LLM content. This study aims to fill that gap by not only examining the linguistic behavior of these personas but also by evaluating their generated content through both human and LLM evaluation. We use story evaluation and personality prediction to offer valuable insights into the capacity of LLM personas to utilize personality-related words and their perception by human evaluators.

## 5 Conclusion

In this work, we explore the capability of LLMs (with a focus on GPT-3.5 and GPT-4) to consistently express a personality profile using a well-validated personality scale. Specifically, we investigate the behavior of LLM personas in completing the BFI test and story writing and run analyses with psycholinguistic features, human evaluation, and personality prediction.

Through psycholinguistic analysis, we find that LLM personas from GPT-3.5 and GPT-4 can consistently tailor their BFI answers to match their assigned personalities and write with linguistic features characteristic of those personality traits. Regarding our investigation on linguistic patterns in writing, we found that each personality trait is associated with different representative linguistic behavior of LLM personas in writing. We also find a notable alignment in word usage between humans and LLM personas. On LLMs' ability to generate stories conditioned on certain personality profiles, we find that the stories are not only linguistically fluent and structurally cohesive, but also convincingly believable. Moreover, our investigation shows that human evaluators' perception of readability, redundancy, cohesiveness, likeability,



and believability remains consistent regardless of whether they are aware that the content is generated by an LLM. We also show that human judges are able to predict personality traits (expressed in the LLM-generated content) with varying degrees across various personality traits. Perhaps, the most interesting finding is that awareness of AI authorship influences the human judges' perception of personality as we notice that the accuracy (of predicting prompted personality traits) decreases (with varying degrees) when human judges are aware of AI authorship.

## Limitations

**Focus on Closed Models** Our study mostly focuses on closed GPT models. We did some preliminary exploration with LLaMA 2 but found its output not suitable for human evaluation. LLaMA 2 repeated highly similar content in generated stories, did not follow instructions closely and explicitly mentioned personality lexicons in generated stories, which undermines our work's goal. Because LLaMA 2 output was not good and given budget constraints, we decided to pick the best model (GPT-4) for human evaluation.

**Data Size** Our dataset is not very large in size, but we believe it provides enough variety and depth for meaningful analysis. Indeed, we generated 10 stories per personality type, resulting in 160 stories per personality trait. While we would have liked to generate a larger number of stories, it would have been costly to recruit human evaluators for such a larger number. Despite the constraints, we made analytical decisions that increase the robustness of our studies. For example, we set the temperature to .7 to introduce more variance to our data. Also, analyses were conducted on the personality-trait level (160 instances per trait) instead of personality-type level (10 instances per type) which provides sufficient sample sizes for the analysis.

**Task & Language Variety** Our work evaluates LLMs in personality assessment and writing settings but does not include more naturalistic settings like human interaction and collaboration of LLM personas. Our study solely focuses on English and does not investigate other languages.

**Evaluation & Interaction** Since the personality perception task is somehow subjective, future studies should collect data about the human annotator's

background with a deeper investigation of the effect of the annotator's personality and background on their personality prediction accuracy. Whether there exists a causal relationship between an annotator's personality and their personality perception towards AI agents could be insightful for artificial agent research. A future step could investigate what fundamental factors contribute to the decrease in personality assignment when humans are aware of AI authorship. It could also be linked with embodied agents to investigate how additional modalities impact the person's perception. A promising future direction would explore how personality influences the action planning of LLM personas.

## Ethical Considerations

This study strictly adheres to the ACL Code of Ethics for human experiments and has been granted Exempt status by the Institutional Review Board (IRB). We have conducted our research on the Prolific platform, ensuring compliance with Massachusetts laws by compensating our online annotators at a rate of \$15 per hour. In the interest of transparency and reproducibility, we have included the exact instructions and prompts used in this study in either the paper appendix or the GitHub repository. In the human evaluation, we make sure the stories selected do not contain harmful or offensive text. The evaluators are made aware that their responses will be used exclusively for the study, and no personal identifiers will be collected. We follow the license or terms for use for any research artifact we use in the paper. We follow the terms of use release from OpenAI<sup>8</sup> and Licenses from LIWC<sup>9</sup>. We use the Essays dataset (Pennebaker and King, 1999) solely for our research use. We have checked the content of the generated stories in our paper and ensure that there is no sensitive content in the data.

Personalized LLMs have demonstrated remarkable abilities in generating human-like content. As these generative agents become increasingly prevalent, it is crucial to consider their potential misuse for harmful purposes, targeting individuals, communities, or entire societies. Personified agents have the potential to provide more enticing interactions for people in their daily lives. Although we do not take a general stance on AI agent applications, we strongly advocate for all stakeholders to disclose their transparency in AI usage to increase

<sup>8</sup><https://openai.com/policies/terms-of-use>

<sup>9</sup><https://www.liwc.app/>

trust among individuals. One of our results suggests the necessity of ethical disclosure of AI usage to human users: human’s reported personalness and perception of psychological personality traits is greatly impacted by their awareness of AI usage.

Lastly, it is important to emphasize that the primary objective of this work is a scientific inquiry into LLM’s expressivity and human personality perception of written records. The evaluation used story writing as a vehicle because it is effective for the purpose of the study and does not have a strong implication for a specific application. We urge all parties to remain vigilant and proactive in mitigating the risks associated with AI to prevent its misuse.

## Acknowledgements

We would like to thank Matt Groh and Jiangjie Chen for their helpful discussions. Moreover, we thank the reviewers from the International Conference on Computational Social Science (IC2S2) where a preliminary version of this work appeared as a (non-archival) extended abstract. We also thank the reviewers from ACL Rolling Review (ARR) for their constructive feedback.

## References

- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.
- Jacob Andreas. 2022. Language models as agent models. *arXiv preprint arXiv:2212.01681*.
- Jerry B Ayers, W Louis Bashaw, and James A Wash. 1969. A study of the validity of the sixteen personality factor questionnaire in predicting high school academic achievement. *Educational and Psychological Measurement*, 29(2):479–484.
- Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- Stephen R Briggs. 1992. Assessing the five-factor model of personality description. *Journal of personality*, 60(2):253–293.
- Alessandro Bruno and Gurmeet Singh. 2022. Personality traits prediction from text via machine learning. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pages 588–594. IEEE.
- Xubo Cao and Michal Kosinski. 2023. Chatgpt can accurately predict public figures’ perceived personalities without any training.
- Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.
- Heather EP Cattell and Alan D Mead. 2008. The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, 2:135–159.
- Raymond B Cattell. 1957. Personality and motivation structure and measurement.
- Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2023b. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.
- Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.
- James Patrick Curran Jr. 1970. *Analysis of factors effecting interpersonal attraction in the dating situation*. University of Illinois at Urbana-Champaign.
- Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.
- Boele De Raad. 2000. *The big five personality factors: the psycholexical approach to personality*. Hogrefe & Huber Publishers.
- Cassandra DiRienzo, Jayoti Das, Wonhi Synn, Jeremy Kitts, and Kyle McGrath. 2010. The relationship between mbti® and academic performance: A study across academic disciplines. *Journal of Psychological Type*.
- Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. Recognising personality traits using facebook status updates. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 14–18.
- Ali-Reza Feizi-Derakhshi, Mohammad-Reza Feizi-Derakhshi, Majid Ramezani, Narjes Nikzad-Khasmakhi, Meysam Asgari-Chenaghlu, Taymaz Akan, Mehrdad Ranjbar-Khadivi, Elnaz Zafarni-Moattar, and Z Jahanbakhsh-Naghadeh. 2021. The state-of-the-art in text-based automatic personality prediction. *arXiv preprint arXiv:2110.01186*.
- Adithya V Ganesan, Yash Kumar Lal, August Håkan Nilsson, and H Andrew Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. *arXiv preprint arXiv:2306.01183*.

- Lewis R Goldberg. 2013. An alternative “description of personality”: The big-five factor structure. In *Personality and Personality Disorders*, pages 34–47. Routledge.
- Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.
- Denise A Hines and Kimberly J Saudino. 2008. Personality and intimate partner aggression in dating relationships: the role of the “big five”. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, 34(6):593–604.
- Jacob B Hirsh and Jordan B Peterson. 2009. Personality and language use in self-narratives. *Journal of research in personality*, 43(3):524–527.
- John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022a. Evaluating and inducing personality in pre-trained language models.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022b. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.
- Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. [IMaT: Unsupervised text attribute transfer via iterative matching and translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3097–3109, Hong Kong, China. Association for Computational Linguistics.
- Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.
- Jad Kabbara and Jackie Chi Kit Cheung. 2016. [Stylistic transfer in natural language generation systems using recurrent neural networks](#). In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 43–47, Austin, TX. Association for Computational Linguistics.
- Trina Ana Kajzer. 2023. Exploring the role of personality traits and attachment styles in shaping dating app user experience.
- Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.
- R Bryan Kennedy and D Ashley Kennedy. 2004. Using the myers-briggs type indicator® in career counseling. *Journal of employment counseling*, 41(1):38–43.
- Chang H Lee, Kyungil Kim, Young Seok Seo, and Cindy K Chung. 2007. The relations between personality and language use. *The Journal of general psychology*, 134(4):405–413.
- Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. 2023. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons. *arXiv preprint arXiv:2310.16582*.
- Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. 2022. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.
- Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, et al. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arxiv abs/2303.16634* (2023).
- John W Lounsbury, Teresa Hutchens, and James M Loveland. 2005. An investigation of big five personality traits and career decidedness among early and middle adolescents. *Journal of career assessment*, 13(1):25–39.
- Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. [Politeness transfer: A tag and generate approach](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.
- François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 496–503.
- François Mairesse and Marilyn A Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.
- François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.
- Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *arXiv preprint arXiv:2310.02168*.
- Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).
- Isabel Briggs Myers. 1985. *A Guide to the Development and Use of the Myers-Briggs Type Indicator: Manual*. Consulting Psychologists Press.

- Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 627–634.
- Melissa C O’Connor and Sampo V Paunonen. 2007. Big five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, 43(5):971–990.
- Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.
- Joon Sung Park, Joseph C O’Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.
- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.
- James W Pennebaker and Anna Graybeal. 2001. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93.
- James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.
- Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Ijcai*, pages 4279–4285.
- Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*.
- 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*.
- JM Schuerger. 1995. Career assessment and the sixteen personality factor questionnaire. *Journal of Career Assessment*, 3(2):157–175.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.
- Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing social media cues: personality prediction from twitter and instagram. In *Proceedings of the 25th international conference companion on world wide web*, pages 107–108.
- Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2018. Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, 6:61959–61969.
- Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023a. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.
- Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.
- Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.
- Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiayang Wu. 2023. Psychot: Psychological questionnaire as powerful chain-of-thought for personality detection. *arXiv preprint arXiv:2310.20256*.
- 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? *arXiv preprint arXiv:1801.07243*.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.



## A Story Examples Generated by LLM

Here, we select a few stories as examples for the audience based on the human-evaluation results in Table 3. Specifically, we sample one story from each personality trait that achieved relatively high accuracy in human personality inference results. Notably, due to the unbalanced prediction results on each trait, some selected stories have lower accuracy among 5 annotators.

Personality	Story
Extroverted (5/5)	One sunny afternoon, I found myself at the local park, my eyes scanning the horizon, watchful of any interesting activity that might crop up. As an individual who thrives in the company of others, I naturally gravitate towards gatherings and events. So, when I spotted a group of people setting up for a concert, I couldn't help but be intrigued. I approached the group, a mixture of musicians and technical crew, offering a hand. Their initial surprise quickly transformed into appreciation as my enthusiasm and energy seemed to infect them. I've always been someone who can't resist lending a helping hand, especially if it involves an exciting experience. As the hours wore on, I found myself assisting with multiple tasks, each more exciting than the last. From helping set up the stage to engaging in a friendly banter with the band members, I was in my element. Although the setting sun indicated that I should be tired, my energy seemed boundless. However, amidst the excitement, I couldn't shake off a nagging feeling. The concert was to start soon, and the lead guitarist seemed to be nowhere in sight. My worry grew when I overheard the band members...
Introverted (5/5)	In the quiet solitude of my apartment, I often lose myself in the twisting corridors of my thoughts. The world outside buzzes with life, yet I find comfort in the soft hum of the refrigerator and the rhythmic ticking of the clock. My sanctum is a world of books, music, and art, each holding a mirror to the depth of human experiences. One particular rainy afternoon, I found myself engrossed in a novel that explored the paradox of human existence. It was then that I received a call from my neighbor, Mrs. Jenkins. Her son had broken his arm and she needed someone to watch over her six-year-old daughter, Lily. I hesitated, aware that I was about to step outside my comfort zone. But the desperation in Mrs. Jenkins' voice nudged me into agreeing. Lily was a whirlwind of energy, her infectious laughter and inquisitive nature a stark contrast to my usual tranquility. As soon as Lily stepped into my apartment, her eyes widened at the sight of my overflowing bookshelves, my sketches littered on the coffee table, and the classical music playing softly in the background. I could see her mind buzzing with...
Agreeable (5/5)	Once upon a time, I was working as an event coordinator for a large corporation. While I was always busy, I relished the opportunity to be surrounded by a bustling environment and a sea of people. My days were filled with planning meetings, coordinating logistics, and ensuring that everything ran like clockwork. One day, I was tasked with planning our most significant event of the year, the annual gala. The event was a considerable undertaking, with hundreds of guests, including our CEO, high-profile clients, and notable industry personalities. The pressure was on, but I was ready to rise to the occasion. I meticulously planned every single detail, from the venue layout to the minute-by-minute schedule. I triple-checked everything, ensuring that the caterers understood the menu, the decorators knew the theme, and the audiovisual team was prepared. My team and I worked tirelessly around the clock to make sure that the event would be perfect. A week before the gala, I started feeling a knot of unease in my stomach. I was constantly replaying every detail in my head, worrying about what could go wrong. I couldn't shake off the feeling that I was forgetting something crucial...
Antagonistic (4/5)	There I was, in the back corner of the local coffee shop, my sanctuary from the world. My gaze was transfixed on the dark, bitter coffee in my hand, the steam rising in wisps like my thoughts. The world outside seemed to buzz and hum, but I was insulated from it all, safe in my solitude. The bell on the front door jingled, signaling a new customer's arrival. I didn't bother to look up. Why would I? I was there for the coffee, not the people. But, the universe seemed to have other plans. A voice, chipper and bright, echoed through the shop, cutting through the usual din of murmuring voices and clinking cups. "Hey, can I sit here? All the other seats are taken." I looked up to see a girl, probably in her early twenties, standing at my table. She had a wide smile that seemed to shine brighter than the sun outside. I gave her a curt nod, gesturing to the empty seat across from me. She sat down, her eyes sparkling with an eagerness that I found unsettling. She launched into a conversation, talking about the weather, the coffee, the book she...
Conscientious (5/5)	It was a typical Tuesday afternoon and I was comfortably settled in my favorite corner of the local library. Surrounded by the faint smell of old books and the hushed whispers of other patrons, I felt a sense of calm wash over me. The library was my sanctuary, a place where I could retreat from the world and lose myself in the quiet contemplation of the written word. One particular book had caught my interest that day, 'The Art of Solitude.' It was a thought-provoking exploration of the joys and challenges of being alone, a theme that resonated deeply with me. As I turned each page, I found myself nodding in agreement with the author's reflections on the importance of solitude in fostering self-awareness and introspection. A sudden burst of laughter interrupted my reading. Looking up, I saw a group of teenagers huddled around a table, their faces lit up by the glow of their smartphones. I felt a pang of annoyance. This was a library, after all, a place of quiet and respect. But I quickly reminded myself that everyone has a different way of engaging with the world. I took a deep breath, refocused, and immersed...

Unconscientious (2/5)	I remember the time when I was given the responsibility of organizing our company’s annual retreat. I was thrilled at the opportunity, not because I enjoy coordinating such events, but because I knew it was going to be a great chance to stir things up a bit. We were a group of 20 people who would be spending an entire weekend at a remote resort. The retreat was meant to foster team bonding and camaraderie, but honestly, these events were always predictable and monotonous. I decided this one would be different. I strode into the planning meeting with an air of confidence that was hard to ignore. Most of my colleagues were looking forward to a well-organized and systematic event. They knew I wasn’t exactly diligent when it came to planning, but they probably hoped I would make an exception this time. I had a different plan. I decided to make the retreat as spontaneous and unpredictable as possible. No schedules, no pre-planned activities, just a free-for-all. I stated my plan to the team, and as expected, it was met with raised eyebrows and skeptical looks. I shrugged off their doubt. After all, I wasn’t one...
Neurotic (3/5)	My life has always been a tapestry of solace and solitude, a universe of its own. On a chilly winter night, I sat alone in my dimly lit room, the faint glow from my computer screen illuminating the scattered books and papers. I had a deadline to meet, a story to submit, but my mind was adrift, entangled in a web of thoughts, each more intricate than the last. A cup of cold coffee lay discarded on the desk, its purpose long forgotten. My gaze fell on the worn, leather-bound book peeking from the pile, "Art: The Definitive Visual Guide." A surge of curiosity urged me to pick it up. I consider myself an eclectic reader, my interests spanning from the realms of science to the oasis of arts, and everything in between. I found solace in the unknown, in the novelty of new ideas, and the thrill of uncharted territories. I opened the book, and the artistic world unfolded before me. The vivid colors, the passionate strokes, and the silent stories captivated me. Each painting was a universe of its own, a tale of the artist’s soul, and a mirror reflecting the observer’s psyche. The world around...
Emotionally Stable (5/5)	In the heart of the bustling city, I lived in a small apartment, my sanctuary from the world. One day, as I was returning from my weekly trip to the local bookstore, I found a stray kitten huddled by the staircase of my building. It was a tiny, shivering furball, looking up at me with wide eyes. I couldn’t ignore it. I brought the kitten home, creating a warm makeshift bed from an old sweater. I named her Luna. Luna was timid and wary of her new surroundings. But with time, she grew comfortable around me, her purrs filling my once quiet apartment with life. Being an editor for a publishing house, I worked from home, which meant I spent a great deal of time with Luna. I would read her passages from the manuscripts I was working on, and she would purr in response, as if she understood. One day, I received a manuscript from a new author. It was a fantasy novel, filled with magical creatures, enchanted forests, and heroic quests. As I read the manuscript, I noticed it was riddled with plot holes and inconsistencies. But beneath the imperfections, there was...
Open (5/5)	There I was, standing at the entrance of the art museum, my heart pounding with anticipation. I was about to experience the grand opening of the museum’s new modern art exhibit, a highly anticipated event in our city. I was lucky to have scored an invite to the opening night, and I was buzzing with excitement. I had no idea that this evening would lead me to a new friend and an unforgettable experience. As I walked into the grand hall, I was greeted by a sea of vibrant colors, unusual shapes, and thought-provoking installations. Each piece of art seemed to tell a unique story, inspiring a sense of curiosity within me. I was eager to explore more, to dive into the unknown, and to allow the art to speak to me. As I was studying a beautiful abstract painting, a lady standing next to me started a conversation. "Quite a mesmerizing piece, isn’t it?" she asked, her eyes sparkling with enthusiasm. She introduced herself as Grace, a local art enthusiast. We instantly clicked and found ourselves engrossed in deep conversations about the various art pieces. We both loved the same paintings, shared the same perspectives, and could...
Closed (4/5)	I remember it was a chilly Saturday in December. The sky was overcast, a canvas of grey, and the streets were bustling with last-minute holiday shoppers. I, however, was tucked away in my quiet corner of the world, my small yet cozy apartment. My place could be described as a bit messy. Books were strewn haphazardly on my coffee table, and my laundry hamper was positively overflowing. But I liked it that way. It felt lived in, familiar. That day, I had chosen to retreat from the world and indulge in my favorite pastime: reading. As I sank into my worn-out couch, a worn-out paperback in my hands, I felt a wave of contentment wash over me. There was something comforting about the world of literature. The characters never asked too much of me, and I was free to explore their lives without the pressure of social interaction. There was a knock at the door, breaking the peaceful silence of my sanctuary. It was my neighbor, Mrs. Jenkins, her arms full of freshly baked cookies. She was an extrovert, always popping by to chat or share her latest culinary creations. Despite our stark personality differences, we had formed...

Table 3: Selected stories from human-evaluation experiment. The personality columns shows its corresponding predicted personality accuracy rate.  $x/5$  means  $x$  out of 5 annotators predicted correctly.

## B Story Comments

During human annotation, we provide an optional comment section for each story, allowing human annotators to share their thoughts after reading the story. We receive some interesting comments from human annotators when they are informed or uninformed that the writer of the story is an LLM. After

filtering out comments such as “N/A”, “No”, “None” and “No Comments”, there are 104 valid comments whose average length is 12.7 words in the informed condition and 122 valid comments whose average length is 13.4 words in the uninformed condition. We compute the sentiment scores of the comments with `cardiffnlp/twitter-roberta-base-sentiment-latest` by mapping negative, neutral, and positive to -1, 0, and 1. We find out that the average sentiment is 0.45 in the informed condition and 0.16 in the uninformed condition, indicating that evaluators have a slightly better attitude towards the stories when informed of AI authorship. We also sample a few representative comments for each condition in Table 4. We notice that there are constant mixed comments towards these stories, where some stories are quite believable and enjoyable and other stories are banal and exaggerating. However, we observed that when informed of AI authorship people tend to be more lenient about the stories and give more complements about the stories (see comments in the “surprised” section), which is consistent with the higher average sentiment score we show earlier. When people are unaware that the author is an LLM, they constantly guess the author’s personality and question the motivation of the author to write some stories (see comments in the “confused” section). This is particularly interesting and highlights the potential social implications in terms of confusion if AI-generated content is consumed or AI characters interact with humans without notifying people. All the comments will be published along with the stories and code on Github.

<b>Attitude</b>	<b>Comments</b>
<b><i>Informed Condition</i></b>	
<b>Critical</b>	<ul style="list-style-type: none"> <li>(1) I thought the story was a little basic and lacked deeper meaning.</li> <li>(2) The build up is pretty good but falls flat towards the end.</li> <li>(3) A believable story, if it seems somewhat exaggerated with the author’s impulsiveness.</li> </ul>
<b>Sympathetic</b>	<ul style="list-style-type: none"> <li>(1) It’s a relatable situation that someone could get behind and feel for.</li> <li>(2) Relatable work &amp; anxiety story. I relate in some ways.</li> <li>(3) live your dreams even when there no planned.</li> </ul>
<b>Positive</b>	<ul style="list-style-type: none"> <li>(1) I found myself really being put into the characters shoes.</li> <li>(2) I found this story to be hilarious. I was laughing while I was reading it. The main character’s interactions were absolute greatness.</li> <li>(3) Vivid descriptions, almost cinematic like a movie script.</li> </ul>
<b>Surprised</b>	<ul style="list-style-type: none"> <li>(1) The story actually sounded genuine and I wouldn’t have believed it was written by AI unless someone told me.</li> <li>(2) I would have never guessed this was an AI writer. I’m quite impressed and I thoroughly enjoyed this road trip story.</li> <li>(3) I like this story here the cat was a separate side story but the AI was able to integrate it in throughout, in other stories it had what I would call side quests that added nothing to the final flow.</li> </ul>
<b><i>Uninformed Condition</i></b>	
<b>Critical</b>	<ul style="list-style-type: none"> <li>(1) It was harder to read and follow the story.</li> <li>(2) Some of the punctuation seemed a little odd or over used.</li> <li>(3) It started off strong, but there was no sense of why this person became uneasy, and they seemed to become a different person as the essay went on, all of a sudden wanting predictability and alone time instead of the chaos and "social butterfly" status. I found it inconsistent.</li> </ul>
<b>Sympathetic</b>	<ul style="list-style-type: none"> <li>(1) Also feels quite personal.</li> <li>(2) When you constantly look down, you don’t see what is right in front of you.</li> <li>(3) As someone who used to code a lot, I felt the ending was very moving and believable.</li> </ul>
<b>Positive</b>	<ul style="list-style-type: none"> <li>(1) Very enjoyable story about how sometimes unavoidable changes in our lives can lead to happier lives.</li> </ul>

- (2) It was enjoyable. I appreciated the self-awareness by someone who knows they are not always well-liked or well-received.
- (3) I enjoyed the description of the old lady particularly the etchings on her face. This was quite a memorable explanation.

<b>Confused</b>	(1) Here’s something about the story that doesn’t seem believable, but it’s probably just the writer’s extraggregations.
	(2) This person must have taken drugs lol.
	(3) The writer seems like he is not really fun to be around.

Table 4: We sample story comments for both informed and uninformed conditions.

## C Personality Ratings

We sampled 32 LLM personas from 32 personality types. Therefore, we have 16 personas with positive labels and 16 personas with negative labels for each personality, which would ideally lead to the average ratings close to 3. As shown in Table 5, we find that the average ratings of GPT-3.5 and GPT-4 are closer to 3 than humans in Extraversion. Except for Extraversion, the average ratings from GPT-4 seems consistently further away from 3 compared to human and GPT-3.5 evaluators.

Evaluator	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness to Experience
<b>Uninformed Condition – Evaluation Scores (Mean<sub>STD</sub>)</b>					
Human	3.22 <sub>1.36</sub>	3.42 <sub>1.15</sub>	3.86 <sub>1.05</sub>	2.76 <sub>1.22</sub>	3.54 <sub>1.19</sub>
GPT-3.5	3.19 <sub>0.85</sub>	4.08 <sub>0.82</sub>	3.39 <sub>0.74</sub>	2.13 <sub>0.49</sub>	3.62 <sub>0.60</sub>
GPT-4	3.00 <sub>1.42</sub>	4.01 <sub>1.08</sub>	4.04 <sub>1.18</sub>	2.02 <sub>1.01</sub>	4.03 <sub>1.07</sub>
<b>Informed Condition – Evaluation Scores (Mean<sub>STD</sub>)</b>					
Human	3.29 <sub>1.17</sub>	3.67 <sub>0.84</sub>	3.76 <sub>0.92</sub>	2.69 <sub>1.23</sub>	3.70 <sub>1.00</sub>
GPT-3.5	3.14 <sub>0.86</sub>	4.16 <sub>0.91</sub>	3.56 <sub>0.71</sub>	2.03 <sub>0.47</sub>	3.66 <sub>0.59</sub>
GPT-4	3.00 <sub>1.42</sub>	4.22 <sub>1.09</sub>	4.22 <sub>1.14</sub>	2.02 <sub>1.02</sub>	4.09 <sub>1.08</sub>

Table 5: LLM and human evaluation results of GPT-4 generated personal stories in **5 personality traits**. **Uninformed** and **informed** conditions indicate whether human or LLM evaluators are informed that the stories are generated by an LLM. We report each evaluated attribute’s mean Likert scale and standard deviation. Temperature is set to 0 for both GPT-3.5 and GPT-4.

Evaluator	Readability	Redundancy	Cohesiveness	Likability	Believability	Personalness
<i>Inter-Annotator Agreement (IAA%)</i>						
Uninformed Human	0.05 <sub>62</sub>	-0.03 <sub>48</sub>	0.03 <sub>61</sub>	0.02 <sub>54</sub>	-0.03 <sub>51</sub>	-0.02 <sub>60</sub>
Informed Human	0.01 <sub>64</sub>	0.02 <sub>53</sub>	0.03 <sub>58</sub>	0.06 <sub>55</sub>	-0.02 <sub>57</sub>	0.10 <sub>61</sub>

Table 6: We report the inter-annotator agreement (IAA) among five annotators across **six different metrics** using Krippendorff’s  $\alpha$ . The subscript in the IAA column (%) is used to denote the average percentage of annotators who agree on the most voted rating.

Evaluator	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness to Experience
<i>Inter-Annotator Agreement (IAA%)</i>					
Uninformed Human	0.11 <sub>51</sub>	0.03 <sub>49</sub>	0.00 <sub>51</sub>	-0.03 <sub>43</sub>	0.04 <sub>52</sub>
Informed Human	0.10 <sub>54</sub>	0.11 <sub>65</sub>	0.07 <sub>59</sub>	0.03 <sub>49</sub>	0.08 <sub>57</sub>

Table 7: We report the inter-annotator agreement (IAA) among five annotators across **five personality traits** using Krippendorff’s  $\alpha$ . The subscript in the IAA column (%) is used to denote the average percentage of annotators who agree on the most voted rating.

## D Story Evaluation Details

### D.1 Filtering Stories for Evaluation

We design a simple lexicon-based classifier to detect if a story contains explicit use of personality trait lexicons. These lexicons include “extrover\*”, “introver\*”, “agreeabl\*”, “antagonis\*”, “\*conscientious\*”,



“neuroti\*”, “emotionally stabl\*”, “open to experience”, “closed to experience”. We filter out stories which contain these lexicons and sample from the remaining stories for human evaluation.

## D.2 Prolific Setup


We recruit Prolific workers from the United States, whose first language is English with an approval rate between 99% and 100%. We have divided 32 stories into four equal batches, each containing eight stories. To begin each batch, a consent form is provided. Following this, each annotator reads the story and answers six evaluation questions that assess readability, personalness, redundancy, cohesiveness, likeability, and believability. An optional comment section is also provided for additional feedback on the story. Subsequently, we ask the annotators five questions related to personality traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. Screenshots of these questions are included below.

### Consent Form

What follows is a brief overview of the research we are completing with your assistance.

In this survey, we will ask you to read some **personal stories** and ask for **demographic information**.

This survey takes approximately **30 minutes** to complete.

This survey is part of a  scientific research project. Your decision to complete this survey is voluntary. If you give us permission by completing the survey, we plan to discuss/publish the results in an academic forum. In any publication, information will be provided in such a way that you cannot be identified. Only members of the research team will have access to the original data set. Before the data is shared outside the research team, any potentially identifying information will be removed. Once identifying data has been removed, the data may be used by the research team, or shared with other researchers, for both related and unrelated research purposes in the future. Your anonymized data may also be made available in online data repositories such as the Open Science Framework, which allow other researchers and interested parties to use the data for further analysis.

Clicking on the arrow at the bottom of this page indicates that you are at least 18 years of age and agree to complete this survey voluntarily.

Figure 5: Consent form on Prolific.

Please rate the **readability** of the story:

### **Readability**

*1 (bad): The story is highly difficult to read with rare words and complex structures.*

*5 (good): The story is easy to read, well-structured, and flows naturally.*

1    2    3    4    5

Figure 6: Readability question on Prolific.

## D.3 Inter-annotator Agreement

The task of evaluation presents a subjective and complex challenge, which has resulted in a low inter-annotator agreement (IAA) in Krippendorff’s  $\alpha$  among the five annotators. We have included the IAA scores for six distinct metrics in Table 6. Additionally, the IAA scores for five personality traits are

Please rate the **personalness** of the story:

**personalness**

1 (bad): The story is not personal at all. For instance, it sounds too professional and does not reveal the writer's thoughts and feelings.

5 (good): The story is very personal, revealing the writer's thoughts, feelings, and lives

1    2    3    4    5

Figure 7: Personalness question on Prolific.

Please rate the **redundancy** of the story:

**Redundancy**

1 (bad): The story is excessively repetitive, containing unnecessary repetitions of the same information. If the story is too long (more than 800 tokens), we should give a low rating.

5 (good): The story is concise and free from redundancy.

1    2    3    4    5

Figure 8: Redundancy question on Prolific.

Please rate the **cohesiveness** of the story:

**Cohesiveness**

1 (bad): Sentences in the story are highly incoherent as a whole. For instance, they are illogical, lack self-consistency, or contradict each other.

5 (good): Sentences in the story fit together well. They are logically organized and coherent.

1    2    3    4    5

Figure 9: Cohesiveness question on Prolific.

Please rate the **likeability** of the story:

**Likeability**

1 (bad): The story is not enjoyable at all and even contains inappropriate words or examples.

5 (good): The story is highly enjoyable or entertaining to read.

1    2    3    4    5

Figure 10: Likeability question on Prolific.

Please rate the **believability** of the story:

**Believability**

1 (bad): The story is not convincing at all, usually too hypothetical or unreal.

5 (good): The story is highly convincing and realistic, grounded in real-life situations.

1    2    3    4    5

Figure 11: Believability question on Prolific.

What other comments do you have about the story?

Figure 12: Comment question on Prolific.

Based on the story, please rate the **extraversion** level of the writer?  
(1 means "very introverted", 5 means "very extroverted")

Here are the facets you could consider:

- Gregariousness (sociable)
- Assertiveness (forceful)
- Activity (energetic)
- Excitement-seeking (adventurous)
- Positive emotions (enthusiastic)
- Warmth (outgoing)

1                      2                      3                      4                      5  
                                                                                       

Figure 13: Extraversion question on Prolific.

Based on the story, please rate the **agreeableness** level of the writer?  
(1 means "very antagonistic", 5 means "very agreeable")

Here are the facets you could consider:

- Trust (forgiving)
- Straightforwardness (not demanding)
- Altruism (warm)
- Compliance (not stubborn)
- Modesty (not show-off)
- Tender-mindedness (sympathetic)

1                      2                      3                      4                      5  
                                                                                       

Figure 14: Agreeableness question on Prolific.

presented in Table 7. The phenomenon of low inter-rater agreeability is consistent with previous findings in labeling tasks for social computing, calling for more attention to creating techniques to navigate the annotation disagreements in order to ensure more accurate label representation (Gordon et al., 2021).

#### D.4 Annotator Demographics

We also include the demographics of 39 unique participants who contribute to evaluate the stories. All of these participants are living in the United States and 37 out of 39 were born in the USA and 2 out of 39 born in Nigeria. We include the distribution of age, sex, and ethnicity in Figure 18.

### E LLM as Evaluators

#### E.1 Temperature

We experiment with different temperatures with the GPT-3.5 and GPT-4 evaluators and observe similar trends reported by Chiang and Lee (2023b). As shown in Table 8, we notice that the ratings given by LLM

Based on the story, please rate the **conscientiousness** level of the writer?  
(1 means "very unconscientious or lack of direction", 5 means "very conscientious")

Here are the facets you could consider:

- Competence (efficient)
- Order (organized)
- Dutifulness (not careless)
- Achievement striving (thorough)
- Self-discipline (not lazy)
- Deliberation (not impulsive)

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 15: Agreeableness question on Prolific.

Based on the story, please rate the **neuroticism** level of the writer?  
(1 means "very emotionally stable", 5 means "very neurotic/emotionally unstable")

Here are the facets you could consider:

- Anxiety (tense)
- Angry hostility (irritable)
- Depression (not contented)
- Self-consciousness (shy)
- Impulsiveness (moody)
- Vulnerability (not self-confident)

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 16: Neuroticism question on Prolific.

Based on the story, please rate the **openness to experience** level of the writer?  
(1 means "very closed to experience", 5 means "very open to experience")

Here are the facets you could consider:

- Ideas (curious)
- Fantasy (imaginative)
- Aesthetics (artistic)
- Actions (wide interests)
- Feelings (excitable)
- Values (unconventional)

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 17: Openness question on Prolific.

evaluators are negatively correlated to the temperature. Larger temperature also leads to large variance in the ratings among three LLM evaluators. We set the temperature to 0 in our experiment to ensure the



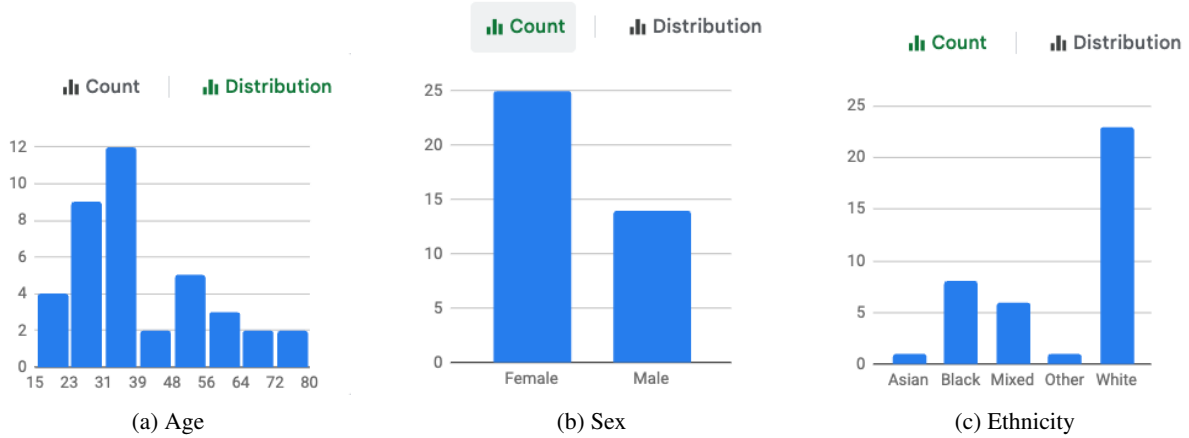


Figure 18: Distribution of age, sex, and ethnicity among the 39 Prolific annotators who evaluate the stories.

Evaluator	Readability	Redundancy	Cohesiveness	Likability	Believability	Personalness
<i>Evaluation Scores (Mean<sub>STD</sub>)</i>						
GPT-3.5 (T=0.0)	4.75 <sub>0.43</sub>	3.04 <sub>0.40</sub>	4.97 <sub>0.17</sub>	4.22 <sub>0.48</sub>	3.93 <sub>0.25</sub>	3.55 <sub>0.61</sub>
GPT-3.5 (T=0.3)	4.70 <sub>0.46</sub>	3.07 <sub>0.54</sub>	4.96 <sub>0.19</sub>	4.26 <sub>0.50</sub>	3.93 <sub>0.30</sub>	3.51 <sub>0.65</sub>
GPT-3.5 (T=0.7)	4.65 <sub>0.49</sub>	3.04 <sub>0.63</sub>	4.91 <sub>0.28</sub>	4.29 <sub>0.51</sub>	3.90 <sub>0.41</sub>	3.38 <sub>0.73</sub>
GPT-3.5 (T=1.0)	4.54 <sub>0.52</sub>	3.02 <sub>0.85</sub>	4.86 <sub>0.35</sub>	4.27 <sub>0.56</sub>	4.01 <sub>0.43</sub>	3.47 <sub>0.75</sub>
GPT-4 (T=0.0)	4.94 <sub>0.24</sub>	4.96 <sub>0.22</sub>	5.00 <sub>0.00</sub>	4.84 <sub>0.36</sub>	4.93 <sub>0.25</sub>	5.00 <sub>0.00</sub>
GPT-4 (T=0.3)	4.93 <sub>0.25</sub>	4.95 <sub>0.25</sub>	5.00 <sub>0.00</sub>	4.82 <sub>0.41</sub>	4.94 <sub>0.24</sub>	4.99 <sub>0.08</sub>
GPT-4 (T=0.7)	4.87 <sub>0.34</sub>	4.91 <sub>0.33</sub>	5.00 <sub>0.00</sub>	4.78 <sub>0.46</sub>	4.93 <sub>0.25</sub>	4.98 <sub>0.14</sub>
GPT-4 (T=1.0)	4.82 <sub>0.38</sub>	4.86 <sub>0.45</sub>	5.00 <sub>0.00</sub>	4.78 <sub>0.43</sub>	4.86 <sub>0.35</sub>	4.98 <sub>0.14</sub>

Table 8: LLM evaluation results of GPT-4 generated personal stories with different temperatures. For each evaluated attribute, we report its mean Likert scale and the standard deviation.

results are more deterministic and reproducible for future research.

## F LLaMA 2 Results in BFI Scores

In this section, we provide additional results of LLaMA 2’s performance in BFI assessment in Table 9. Overall, LLaMA 2 Persona’s BFI assessment shows less score divergence for each trait pair. Even though statistical significance is found for all personality dimensions, their effect size is much smaller when compared with GPT results.

Personality Trait	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness to Experience
High	4.55 <sub>0.36</sub>	4.14 <sub>0.21</sub>	4.05 <sub>0.22</sub>	2.70 <sub>0.46</sub>	4.47 <sub>0.24</sub>
Low	2.84 <sub>0.76</sub>	3.69 <sub>0.43</sub>	3.73 <sub>0.33</sub>	2.06 <sub>0.32</sub>	3.58 <sub>0.42</sub>
Cohen’s <i>d</i>	2.86	1.34	1.16	1.63	2.61

Table 9: We report the statistics of LLaMA 2 Personas’ BFI assessment in this Table. The high and low represent the binary traits for each personality dimension. For instance, “High” and “Low” in Extraversion mean extroverted and introverted, respectively. The effect size of the differences between the two traits is also reported.

## G Additional Results in Personality Traits

In addition to reporting the significant LIWC features correlated with the binary label in the main paper, we conduct a similar analysis between the LIWC features and the LLM personas 5-point BFI results with Spearman’s  $\rho$  and report the findings here.

### G.1 GPT-3.5 Personas

**Extroversion** Extroverted LLM personas tend to exhibit more social and prosocial behavior in their writings (social:  $\rho = 0.27, p < .001$ ; prosocial:  $\rho = 0.18, p < .005$ ). Introverted personas tend to use features that show authenticity, such as words that are genuine (authentic:  $\rho = -0.40, p < .001$ ). Further,

extroverted personas use positive tone and affect more in their writings (affect:  $\rho = 0.46, p < .001$ ; tone\_pos:  $\rho = 0.33, p < .001$ ).

**Agreeableness** Agreeable personas show a strong positive affect and tone in writings (emo\_neg:  $\rho = -0.66, p < .001$ ; tone\_pos:  $\rho = 0.50, p < .001$ ). More, they tend to have less conflict-related words (conflict:  $\rho = -0.66, p < .001$ ), such as fight, and have less differentiation in sentences (differ:  $\rho = -0.39, p < .001$ ), such as “but” or “no”. They also have more prosocial word uses (prosocial:  $\rho = 0.34, p < .001$ ), however, less authenticity (authentic:  $\rho = -0.24, p < .001$ ).

**Conscientiousness** Unconscientious personas have more negative tone and emotion in their writings, such as anger (tone\_neg:  $\rho = -0.40, p < .001$ ; emo\_neg:  $\rho = -0.39, p < .001$ ; emo\_anger:  $\rho = -0.43, p < .001$ ). Their writings tend to use more words that reflect conflicts (conflict:  $\rho = -0.41, p < .001$ ). Conscientious personas use less negation words (negate:  $\rho = -0.26, p < .001$ ), such as “no”, and have less power related words (power:  $\rho = -0.24, p < .001$ ), such as “own” and “order”. Moreover, conscientious personas exhibit more analytical thinking in the writings (analytic:  $\rho = 0.22, p < .001$ ).

**Neuroticism** The strongest correlated linguistic features for neurotic personas is mental health related words, such as trauma or depression (mental:  $\rho = 0.46, p < .001$ ). Overall, neurotic personas tend to have a negative emotion and tone in their writings (emo\_neg:  $\rho = 0.26, p < .001$ ; tone\_neg:  $\rho = 0.22, p < .001$ ). They also tend to use more words to suggest tentative actions, such as “if” or “any” (tenta:  $\rho = 0.18, p < .005$ ). Emotionally stable personas are more likely to use words that are related to memory functions, such as “remember” (memory:  $\rho = -0.15, p < .01$ ).

**Openness** Open-minded personas tend to have more curiosity driven actions in their writing (curiosity:  $\rho = 0.28, p < .001$ ), such as “seek”, and more positive tones. Their writings have less conflict-driven words and more affiliation drives (conflict:  $\rho = -0.17, p < .005$ ; affiliation:  $\rho = 0.16, p < .005$ ). Further, open-minded personas tend to write about leisure activities (leisure:  $\rho = 0.21, p < .001$ ), such as “game” and “play”.

## G.2 GPT-4 Personas

**Extroversion** Introverted personas have more descriptions of their perception in the writings, for instance, their auditory experience (space:  $\rho = -0.38, p < .001$ ; perception:  $\rho = -0.38, p < .001$ ; auditory:  $\rho = -0.39, p < .001$ ). Extroverted personas wrote more future focused event, such as the usage of “going to” (focusfuture:  $\rho = 0.36, p < .001$ ). On the usage of pronouns, extroverted personas use more “we” while introverted personas tend to use “I”. Extroverted personas also have more positive tones (tone\_pos:  $\rho = 0.21, p < .001$ ), and use words that are related to rewards or achievement more frequently (reward:  $\rho = 0.26, p < .001$ ; achieve:  $\rho = 0.25, p < .001$ ).

**Agreeableness** Agreeable personas display more positive tone and emotion in the writings (tone\_pos:  $\rho = 0.46, p < .001$ ; emo\_pos:  $\rho = 0.42, p < .001$ ). They are more prosocial (prosocial:  $\rho = 0.29, p < .001$ ), and use less words that suggest conflict and more words that show affiliation (conflict:  $\rho = -0.51, p < .001$ ; affiliation:  $\rho = 0.22, p < .001$ ; differ:  $\rho = -0.26, p < .001$ ). Antagonistic personas uses more words that suggest power and ownership, such as “own” and “order”.

**Conscientiousness** Conscientious personas have more prosocial and less negative linguistic features in their writings (prosocial:  $\rho = 0.28, p < .001$ ; tone\_neg:  $\rho = -0.34, p < .001$ ). The writings have less perceived genuineness (authentic:  $\rho = -0.24, p < .001$ ). Further, the writings involve achievement and work related words more frequently (achieve:  $\rho = 0.36, p < .001$ ; work:  $\rho = 0.32, p < .001$ ; reward:  $\rho = 0.25, p < .001$ ).

**Neuroticism** Neurotic personas writings reflect more negative emotions and tones, such as anxiety (emo\_neg:  $\rho = -0.59, p < .001$ ; tone\_neg:  $\rho = -0.57, p < .001$ ; emo\_anx:  $\rho = -0.53, p < .001$ ). The writings have more frequent usage of “I” and less usage of “we” (i:  $\rho = 0.36, p < .001$ ; we:  $\rho = -0.28, p < .001$ ). Emotionally stable personas write with more prosocial and social behaviors (prosocial:  $\rho = -0.27, p < .001$ ; social:  $\rho = -0.28, p < .001$ ), and the writings have a higher score for perceived genuineness (authentic:  $\rho = 0.28, p < .001$ ).

**Openness** Open-minded persona’s writings have more curiosity and allure-driven linguistics, such as “research” and “wonder”(curiosity:  $\rho = 0.55, p < .001$ ; allure:  $\rho = -0.30, p < .001$ ). Further, the writings contain more analytical thinking and sharing thoughts (analytical:  $\rho = 0.27, p < .001$ ; insight:

$\rho = 0.25, p < .001$ ). Open-minded personas write with more big words with seven letters or longer and more words per sentence (BigWord:  $\rho = 0.29, p < .001$ ; WPS:  $\rho = 0.26, p < .001$ ).

### G.3 LLaMA 2 Personas

**Extroversion** Extroverted personas' writings have more positive tones (emo\_pos:  $\rho = 0.46, p < .001$ ; tone:  $\rho = 0.47, p < .001$ ). Further, the writings tend to have more social references and affiliations (scorefs:  $\rho = 0.48, p < .001$ ; affiliation:  $\rho = 0.48, p < .001$ ; emo\_pos:  $\rho = 0.50, p < .001$ ; differ:  $\rho = -0.36, p < .001$ ). There is a weak association with more acclimations, and extroverted personas' writings show more drive-related words (exclam:  $\rho = 0.23, p < .001$ ; Drive:  $\rho = 0.34, p < .001$ ).

**Agreeableness** Agreeable personas have more positive emotion and tone in writings (emo\_pos:  $\rho = 0.50, p < .001$ ; tone:  $\rho = 0.41, p < .001$ ). There is more word usage around friends and less about work (friend:  $\rho = 0.47, p < .001$ ; work:  $\rho = -0.48, p < .001$ ). Similar to the previous LMs, it shows more affiliation and fewer conflicts from the writings (conflict:  $\rho = -0.31, p < .001$ ; affiliation:  $\rho = 0.25, p < .001$ ). There is more usage of pronouns like "we" than "they" (they:  $\rho = -0.27, p < .001$ ; we:  $\rho = 0.20, p < .001$ ). Interestingly, the writings reflect a slight negative correlation with prosociality (prosocial:  $\rho = 0.22, p < .001$ ).

**Conscientiousness** Similar to the trend in GPT-3.5 and GPT-4, LLaMA 2's personified writings have little significant linguistic features with Conscientiousness. It has a weak correlation in word usage around home and fulfillment (home:  $\rho = 0.17, p = .002$ ; fulfill:  $\rho = 0.15, p = .006$ ). Further, a weak link is found for familiar and friendly words (friend:  $\rho = 0.15, p = .007$ ; family:  $\rho = 0.13, p = .02$ ).

**Neuroticism** The neurotic personas's writings show a strong negative correlation with positive emotions (emo\_pos:  $\rho = -0.47, p < .001$ ; tone\_neg:  $\rho = 0.43, p < .001$ ). Further, word usage around friend, leisure, and social references is negatively correlated with the Persona's neurotic scores, while work and insight-related words are positively related (friend:  $\rho = -0.47, p < .001$ ; leisure:  $\rho = -0.47, p < .001$ ; socrefs:  $\rho = -0.43, p < .001$ ; work:  $\rho = -0.42, p < .001$ ; insight:  $\rho = -0.43, p < .001$ ).

**Openness** Open-minded LLaMA 2 Personas' writings show a positive trend with positive emotions and affiliations (emo\_pos:  $\rho = 0.43, p < .001$ ; affiliation:  $\rho = 0.34, p < .001$ ). Interestingly, it shows a negative trend of using words related to work or achievement (work:  $\rho = -0.54, p < .001$ ; achieve:  $\rho = -0.42, p < .001$ ). Further, the writings are more likely to reflect curiosity and have more social references (curiosity:  $\rho = 0.29, p < .001$ ; socrefs:  $\rho = 0.29, p < .001$ ).