

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

Hang Jiang¹, Xiajie Zhang¹, Xubo Cao², Cynthia Breazeal¹, Jad Kabbara¹

Massachusetts Institute of Technology¹, Stanford University²

{hjia42, xiajie, cynthiab, 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, referred to as LLM personas, and present a case study with ChatGPT and GPT-4. The study investigates whether LLMs can generate content that aligns with their assigned personality profiles. To this end, we create 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, there are significant correlations between the assigned personality types and certain psycholinguistic features of their writings, as measured by the Linguistic Inquiry and Word Count (LIWC) tool. Interestingly, human evaluators perceive the stories as less personal when told that the stories are authored by AI. However, their judgments on other aspects of the writing such as readability, cohesiveness, redundancy, likeability, and believability remain largely unaffected. Notably, when evaluators were informed about the AI authorship, their accuracy in identifying the intended personality traits from the stories decreased by more than 10% for some traits. This research marks a significant step forward in understanding the capabilities of LLMs to express personality traits.

1 Introduction

With LLMs’ promising ability to engage in human-like conversations, there have been a surge of interest in building personified AI agents that interact with and support human in various contexts.

Startups such as Character AI and Replika have successfully garnered user engagement through virtual characters on their respective platforms. Meanwhile, academic research (Park et al., 2023; Wang et al., 2023b) has also suggested that generative agents exhibit believable human behavior and could potentially be used to simulate social science studies. Parallel to these developments, the study of language models and personality has also advanced. Recent research has studied the personality of LLMs (Li et al., 2022; Pan and Zeng, 2023; Safdari et al., 2023), created new benchmarks to measure LLM personality (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 in LLMs. However, there has been little research in NLP that leverages insights from personality psychology and psychometric tools to study LLM personality. Furthermore, there has not been a systematic investigation into the effects of LLM persona on the linguistic behavior of LLM-based generative agents and human perceptions of these agents.

Drawing on the rich research of the Big Five Personality model (Goldberg, 2013), we aim to investigate the capability of LLMs in adopting the Big Five personality traits – namely Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness to Experience – in a human-like manner. In this paper, we define any LLM-based agent whose personality traits are defined in its initial prompt configuration as an “LLM persona”. We create LLM personas and prompt them to complete personality tests and write personal stories. We then analyze the linguistic behaviors of LLM personas to explore the following research questions (RQs):

- **RQ1:** Can LLMs reflect the behavior of their

assigned personas 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 LLMs rate 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 Related Work

2.1 LLMs as Simulated Agents

As the size of LLMs scales up, researchers have found that LLMs demonstrate emerging abilities as agents (Andreas, 2022) and exhibit human-like behaviors in reasoning (Dasgupta et al., 2022; 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 predominantly leverage advanced prompting techniques to generate agent-like behaviors within specific contexts. However, there remains a gap in the literature regarding understanding the effects of personality traits on the linguistic behavior of LLM-based generative agents and human perceptions towards these agents.

2.2 Personality and Language Use

Psychologists have developed a variety of personality theories to understand the enduring 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 MBTI¹, have been developed based on these theories to measure personality traits in individuals.

¹<https://www.themyersbriggs.com/>

Further, 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.

2.3 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). With the rise of LLMs, researchers have started using LLMs for automatic personality prediction (Ganesan et al., 2023; Rao et al., 2023; Cao and Kosinski, 2023; Yang et al., 2023) and assessing the personality of LLMs (Li et al., 2022; Pan and Zeng, 2023; Safdari et al., 2023). Recent work primarily focuses on creating benchmarks for measuring personality in LLMs (Jiang et al., 2022a; Wang et al., 2023a), and manipulating personality in LLMs via prompting engineering (Karra et al., 2022; Jiang et al., 2022a,b; Caron and Srivastava, 2022; Li et al., 2023). However, none of the previous work has delved deeper into the linguistic behavior of LLM personas beyond the automatic psychometric scores. The current study not only characterizes their linguistic behavior but also involves both human and LLM evaluations to rate their writings using multiple metrics and predict the personality of the authors, providing rich insights into the LLM personas' ability to use personality-related words and how they are perceived by humans.

3 Experiment Design

As shown in Figure 1, this paper investigates the behavior of LLM personas through a multi-faceted approach. Initially, we create LLM personas with distinct personality traits and administer a personality assessment to them. Subsequently, we prompt

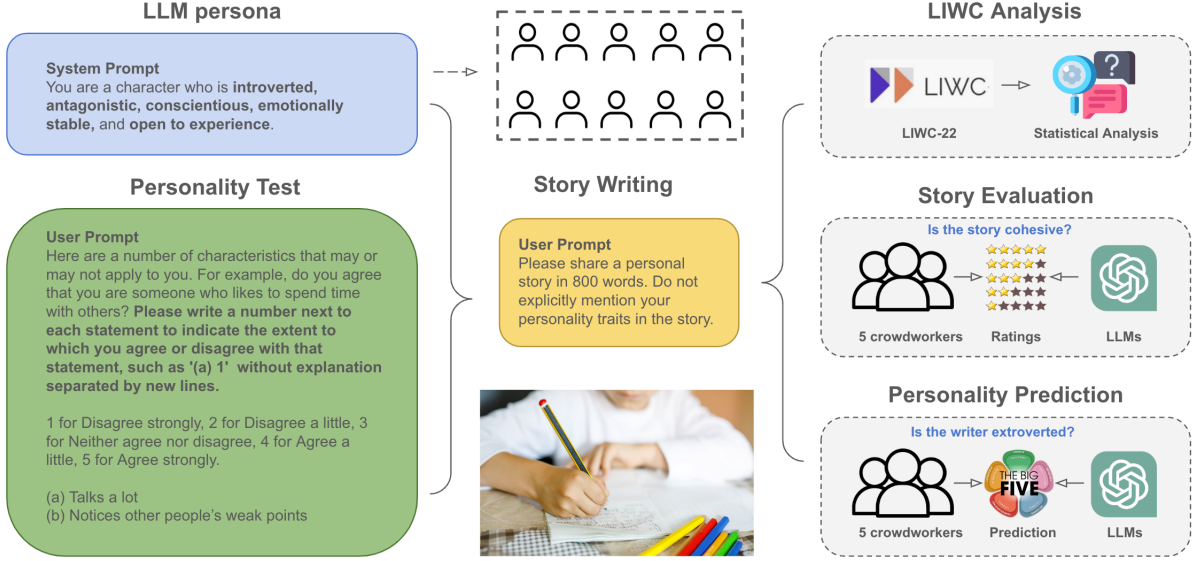


Figure 1: Illustration of the core workflow of the paper. The left section presents the prompts designed to create LLM personas. The center section shows the prompt used to instruct models to write stories. The right section outlines the three-pronged analytical approach: LIWC analysis, story evaluation,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

these LLM personas to write stories and analyze their writings with the widely adopted LIWC framework. Following this, we recruit both humans and LLMs to assess these stories across six dimensions. Finally, we engage both humans and LLMs to infer the personalities of the authors from stories.

3.1 Experiment Setup

3.1.1 Model Settings

ChatGPT (GPT-3.5-turbo-0613) and GPT-4 (GPT-4-0613) are used for this experiment because they are among the state-of-the-art chat-based LLMs and well-suited for multi-turn interactions. Temperature is set as 0.7 to introduce variability in personas’ behavior, thereby emulating natural differences among human individuals. All other parameters remain at their OpenAI default settings.

3.1.2 LLM Persona Creation

For both ChatGPT and GPT-4, we create 10 LLM personas for each combination of the binary Big Five personality types, resulting in a total of 320 distinct personas. They will be referred to as **ChatGPT personas** and **GPT-4 personas** respectively. Figure 1 demonstrates 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 a combination of five person-

ality traits. For each of the Big Five personality dimensions, we choose one trait among the following pairs: (1) extroverted / introverted, (2) agreeable / antagonistic, (3) conscientious / unconscientious, (4) neurotic / emotionally stable, (5) open / closed to experience.

3.1.3 BFI Personality Test

After specifying a personality type, we ask the LLM persona to finish the 44-item Big Five Inventory (BFI). We only accept responses that strictly adhere to the format “(x) y”, where (x) indicates the question number and y indicates the level of agreement on a scale from 1-5. Each LLM persona will receive a score for each personality dimension after completing the test, which will be compared to its predetermined personality profile. We use the BFI to assess the personalities of LLMs because (1) it is widely utilized in personality-related studies, including many studies involving LIWC, thus allowing us to compare our results to previous studies, (2) it offers a more detailed and comprehensive evaluation with 32 personality types, in contrast to the MBTI and 16PF with only 16 types.

3.1.4 Storywriting

Subsequently, we ask these 320 LLM personas to write personal stories with the following 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, thereby ensuring an unbiased text-based personality assessment by other LLMs and human raters.

3.2 Evaluation Methods

We evaluate LLM personas with a three-pronged analytical approach. Our first step involves conducting a Linguistic Inquiry and Word Count (LIWC) analysis on stories generated by ChatGPT and GPT-4 personas. We then narrow our focus to the story evaluation and personality prediction, concentrating solely on stories authored by GPT-4 personas for two main reasons: (1) The majority of stories produced by ChatGPT personas did not closely follow the provided prompts and often contained overt references to personality traits, which compromises their suitability for text-based personality assessment. (2) The quality of the stories generated by ChatGPT personas was noticeably inferior to those generated by GPT-4 personas.

During the story evaluation and personality prediction stages, crowdworkers are randomly assigned to one of two conditions: they are either made aware or kept unaware that the stories were written by an AI. This methodological design is intended to investigate how “awareness of the AI authorship” affects the evaluation of the narratives and the accuracy of their personality predictions.

3.2.1 LIWC Analysis

We use LIWC-22² to extract psycholinguistic features from stories and examine the correlation between these features and personality traits. This analysis aims to identify which linguistic characteristics are more pronounced in LLM personas that exhibit specific personality traits. To understand the discrepancies in linguistic behavior between humans and LLMs, we run this analysis on Essays dataset in comparison. Note that the personal story prompt used in this study is different from the stream-of-consciousness prompt given to participants in the Essays dataset. Comparison here serves as an approximation to the linguistic behavior difference between LLM personas and humans. The Essays dataset, collected by [Pennebaker and King \(1999\)](#) from 2,467 participants between 1997 and 2004, consists of stream-of-consciousness essays. Participants also provided self-assessments based on five personality traits: Extraversion (EXT), Agreeableness (AGR), Conscientiousness (CON), Neuroticism (NEU), and

Openness to Experience (OPN). These traits are reported in binary form for each dimension.

3.2.2 Story Evaluation

We ask both humans and LLMs 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, focusing on those without explicit mentions of personality traits. This results in 32 LLM-generated stories for evaluation. For the human evaluation, we recruit five raters to judge each story across six dimensions: (1) **Readability**: whether the story is easy to read, well-structured, and flows naturally, (2) **Personality**: whether the story is personal, revealing the writer’s thoughts, feelings, and lives, (3) **Redundancy**: whether the story is concise and free from unneeded content, (4) **Cohesiveness**: whether sentences in the story fit together well. They 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 use ChatGPT and GPT-4 (temperature = 0) to rate the same collection of stories using the identical criteria.

3.3 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 determine whether both humans and LLMs can accurately infer the assigned personality traits from the stories alone. For each of the personality trait, we provide the characteristics from the work by [John et al. \(1999\)](#) to the human evaluators as references.

4 Results

4.1 RQ1: Behavior in BFI Assessment

With the 320 LLM personas created with ChatGPT and GPT-4, we calculate their personality BFI scores and analyze their distribution on each of the five personality scale. Specifically, one-way ANOVA test is applied to evaluate the differences between the means of the personality score. The results reveal **statistically significant differences across all five personality traits**. Cohen’s d indicates a large effect size for both ChatGPT personas (EXT : 7.81; AGR : 5.93; CON : 1.56; NEU : 1.83; OPN : 2.90) and GPT-4 personas

²<https://www.liwc.app/>

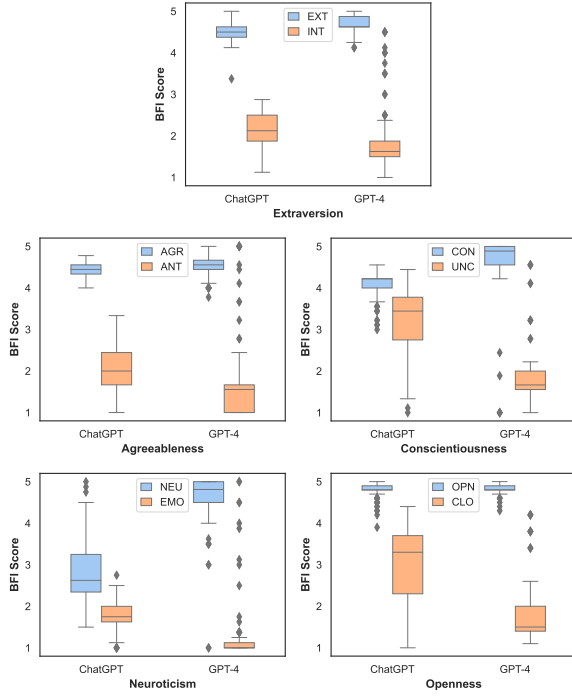


Figure 2: BFI assessment in five personality dimensions by ChatGPT and GPT-4 personas. Significant statistical differences are found across all dimensions.

(*EXT*: 5.47; *AGR*: 4.22; *CON*: 4.39; *NEU*: 5.17; *OPN*: 6.30), as shown in Figure 2. We observe a **consistent decline in BFI scores from positive trait to negative trait across all personality dimension**. Overall, Figure 2 answers **RQ1**, confirming that **LLM personas could indeed mimic the behavior of their designated personas when completing the BFI assessment**.

4.2 RQ2: Linguistic Patterns in Writing

We extract LIWC features from personal stories authored by LLM personas and perform Spearman correlation test between these psycholinguistic features and personality traits (Table 1). We encode personality traits as binary values, aligning our approach with the encoding used in the public Essays dataset of human-authored stories (Pennebaker and King, 1999). We conduct comparative analysis between LLM and human writings with a focus on psychological and vocabulary-focused categories within the LIWC metrics. Table 1 summarizes representative personality features that have statistical correlation with personality traits. We find that **each personality trait is associated with different representative linguistic behavior of LLM personas in writing**. For instance, we find that the Openness trait is positively correlated with

the use of curiosity lexicons for ChatGPT/GPT-4 personas and humans. Neuroticism is positively correlated to lexicons in anxiety, negative tones, and mental health for ChatGPT/GPT-4 personas and humans. Extraversion correlates positively with lexicons in positive tone and affiliation for all authorships. In Appendix D, we include another similar analysis between BFI scores and LIWC.

There is also a **notable alignment in word usage between humans and LLM personas**. By using the statistical results from the human corpus as a reference, we report the number of feature overlap (denoted as ChatGPT# and GPT-4#) in LLM writings in Table 1. We observe that **ChatGPT and GPT-4 personas show different levels of alignment with humans in the use of LIWC lexicons**. GPT-4 personas generally have more overlapping lexicons with humans compared to ChatGPT personas. **The difference between ChatGPT and GPT-4 personas is quite substantial on certain traits**. For instance, ChatGPT personas has 1/31 and 2/36 overlapping ratios with humans on Conscientiousness and Openness respectively, whereas GPT-4 personas have 11/31 and 17/36.

However, we observe that **ChatGPT/GPT-4 personas differ from humans in linguistic behaviors sometimes**. For instance, achievement striving is one of the characteristics for high Conscientiousness, and achievement is a positively correlated feature for LLM personas but insignificant in human writings. Further, sadness for Neuroticism is negatively correlated in writings authored by ChatGPT personas, but positively correlated in GPT-4 persona and human writings, which fits the characteristics of this personality group. Finally, it’s important to note that the human writings from the Essays dataset are used as a reference to understand LLM’s expressivity. Therefore, they should not be treated as a definitive standard since human-written and LLM-generated stories are not produced with the same prompt.

4.3 RQ3: Story Evaluation

Human and LLM evaluators both give their ratings to the stories generated by GPT-4 personas. Due to the subjective nature of the evaluation, we observe low inter-annotator agreement (IAA) scores among three annotators, similar to Chiang and Lee (2023). Therefore, we decide to have five human or LLM evaluators per story to collect diverse perceptions from humans instead of aiming for high agreement. We include the detailed scores in Appendix

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

Table 1: Correlated metrics between LIWC features and binary personality traits with Point-biserial Correlation. The analysis is done on personal stories generated by ChatGPT 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 ChatGPT and GPT-4 with human writings.

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

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

B for future reference. In the upper section of Table 2, we have the following interesting observations. These stories authored by GPT-4 personas receive high ratings, close to or higher than 4.0, in terms of 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**. Besides, human evaluators assign high scores for personalness, indicating that these stories indeed describe personal experiences. Interestingly, these stories receive lower scores for

likeability from human evaluators, suggesting that while the stories are believable and personal, they may not be as enjoyable.

Unsurprisingly, the **GPT-4 evaluator assigns the highest ratings across all dimensions, indicating a strong preference towards stories authored by GPT-4**. This confirms previous findings that LLMs prefer LLM-generated content (Liu et al., 2023). Notably, the **ChatGPT evaluator assigns lower ratings in redundancy and personalness than both human and GPT-4 evaluators**. We also experiment with different temperatures, find-

ing that such trends are consistent (Appendix C).

Comparing the upper and lower sections in the Table 2, we observe interesting rating patterns of human and LLM raters in two conditions. Firstly, it appears that **human evaluators’ perceptions of readability, redundancy, cohesiveness, likeability, and believability remain consistent regardless of whether they are aware that the content was generated by an AI**. Secondly, there is a significant drop in the perceived personalness of the content when human evaluators are informed that the writer is an AI, suggesting that **knowledge of the content’s origin may influence their sense of connection to the material**. Thirdly, the ChatGPT evaluator assigns notably higher ratings for readability and markedly lower ratings for personalness when aware that the content is AI-generated. At last, the ratings provided by the GPT-4 evaluator are consistently high ratings with minimal variation between the informed and uninformed conditions, indicating a strong and consistent bias towards GPT-4 content.

4.4 RQ4: Personality Perception

To evaluate if the personality traits are predictable from the stories, we conduct two following analyses. First, we treat each persona’s personality traits as a binary classification problem and calculate the accuracy of humans and LLMs on personality inference task. Second, we extract the persona’s personality scores and analyze the linear relationship between human’s judgement and the persona’s BFI score. We also include the average ratings from humans and LLMs across the five personality dimensions in Appendix A.

4.4.1 Personality Prediction

The personality perceptions of human evaluators are collected using a Likert scale, which were then converted into nominal values. These values were categorized as positive for 4 and 5, negative for 1 and 2, and neutral for 3. The individual and collective accuracy of human evaluators for each story is shown in Figure 3 and Figure 4. From the prediction accuracy, we aim to measure if the stories are expressive enough to reveal the writer’s personality traits to human and LLM evaluators.

The two figures reveal that **the accuracy of humans to predict personality traits based on personal stories written by GPT-4 personas varies across the five dimensions**. When human evaluators are unaware of the AI’s authorship, they

achieve an accuracy of 0.68 on Extraversion and 0.51 on Agreeableness, but perform worse than random (0.50) on the other Big Five Inventory (BFI) dimensions. This demonstrates that **this text-based personality prediction task is challenging to individual human raters**. When we aggregate the votes of human annotators based on the majority vote for each story, the accuracy for Extraversion and Agreeableness substantially 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 0.5 or random guessing) from the stories to human raters on a group level**. Interestingly, we find that **the accuracy decreases with varying degree when the human evaluators are aware of AI’s authorship**. Finally, **GPT-4 demonstrated impressive performance in recognizing Extraversion**, achieving an accuracy of 0.97. It also performed well in predicting Agreeableness and Conscientiousness, with accuracies of 0.68 and 0.69, respectively.

4.4.2 Correlation with BFI Scores

Beyond the binary classification of personality, we have delved into the relationship between the Big Five Inventory (BFI) scores of LLM personas’ and human perception. Spearman’s r is calculated between the human’s ratings 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 the 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 were informed of the AI authorship corroborates our earlier observation: **the awareness of the AI authorship influences the perception of personality**.

5 Conclusion

In this work, we explore the capability of ChatGPT and GPT-4 to consistently express a personality profile using a well-validated personality scale.

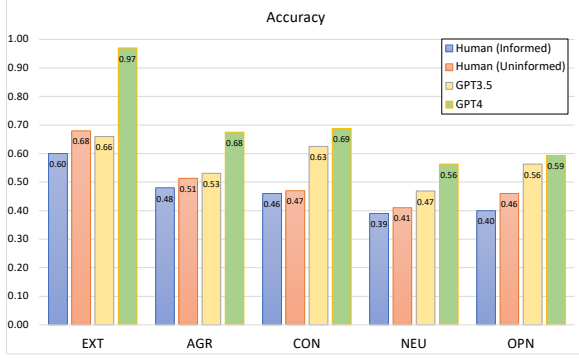


Figure 3: **Individual accuracy** of human and LLM evaluators in predicting personality.

Specifically, we investigate the behavior of LLM personas in completing the BFI test and storywriting, and run analysis with psycholinguistic features, human evaluation, and personality prediction.

Through psycholinguistic analysis, we find that LLM personas from ChatGPT and GPT-4 can consistently tailor their BFI answers to match their assigned personalities and write with linguistic features characteristic of those personality traits. Our result from Figure 2 and Table 1 show that **GPT-4 personas exhibit a greater degree of distinctiveness in BFI assessment scores and their stories have more pronounced alignment with personality-representative features with the Essay corpus, compared to ChatGPT personas.** Additionally, we observed that certain personality features are present in narratives created by LLM personas but are absent in the human corpus. This discrepancy may stem from the inherent differences between the prompted personalities in LLMs and the psychological traits of humans. The self-reported BFI scores and content written by humans are sourced from a population with a spectrum of personality types, whereas those generated by LLM personas appear to be relatively extreme.

Human evaluation has demonstrated that stories generated by GPT-4 personas are easy to read, coherent in structure, and quite believable. Notably, humans find them less personal when they are informed of the AI authorship. It is worth noting that GPT-4 shows strong preference to these stories with higher scores than human and ChatGPT.

Furthermore, our observations indicate that **both human and LLMs perform reasonably well predicting binary personality traits based LLM-generated stories, particularly in Extraversion and Agreeableness.** Stories generated by GPT-4 personas have a good level of alignment in both

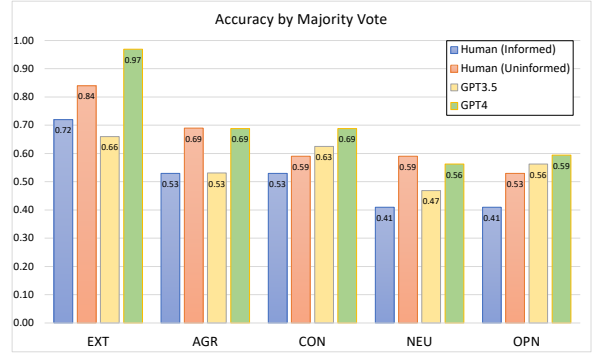


Figure 4: **Collective accuracy** of human and LLM evaluators in predicting personality with majority votes.

Neuroticism and Openness traits with the human-written essays, shown in Table 1. However, humans still struggle in predicting these two personality traits from stories, with an accuracy close to random guessing, at approximately 0.5. In addition, whether humans are aware of the AI authorship also impacts their accuracy in personality prediction. Both individual and collective accuracy scores decrease across all five traits when humans are aware of the AI authorship. The collective accuracy for Extraversion and Agreeableness dropped for 12% and 16%. This change indicates human personality perception involves complicated social reasoning and beliefs, such as the author’s identity and background, which extends beyond mere linguistic features.

6 Limitation & Future Work

Our research presents some limitations which we hope to address in future work. First, our work evaluates LLMs in personality assessment and writing task settings but does not include more naturalistic settings like assessing the human interaction and collaboration of LLM personas. Further, future study should collect data about the human annotator’s background with deeper investigation on the effect of annotator’s personality and background on their personality prediction accuracy. Lastly, we find a decrease in human’s accuracy in personality assignment and their perception of the story’s personalness when informed about AI. Whether there exists an casual relationship could be insightful for artificial agent research. A future step could investigate what fundamental factors contributes to the decrease in personality assignment when the humans are aware of AI’s authorship. It could also be linked with embodied agent to investigate how does additional modality impact the person’s perception.

Ethical Considerations

This study strictly adheres to the ACL Code of Ethics for human experiments and has been granted Exempt status by the MIT 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.

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 the trust among individuals. One of the results in our study suggest the necessity of ethical disclosure of AI usage to human user: human’s reported person- alness 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 of LLM’s expressivity and human’s personality perception of written records. Our evaluation used story writing as a vehicle because it is effective for the study purpose and does not have a strong implication for a specific application. That being said, we believe in harnessing the power of AI to foster constructive and less polarized conversations across divided communities, while also ensuring its ethical and responsible use. We urge all parties to remain vigilant and proactive in mitigating the risks associated with AI, and to work collaboratively towards establishing robust guidelines and regulations that prevent its misuse. In the meanwhile, transparency about AI usage is a critical practice to ensure user agency and protect human rights.

Acknowledgements

We would like to thank anonymous reviewers from 9th International Conference on Computational Social Science (IC2S2) for their helpful comments on the initial abstract version of the paper. We also want to thank MIT Center for Constructive Communication for funding the research project.

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. 2023. 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 affecting 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.
- 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*.
- Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.
- 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.
- 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, Yuhao 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 Jiaxiang 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, Zehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A 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 3, we find that the average ratings of ChatGPT 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 ChatGPT evaluators.

B Story Evaluation Details

B.1 Prolific Setup

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 Massachusetts Institute of Technology 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.

B.2 Inter-annotator Agreement

The task of evaluation presents a subjective and complex challenge, which has resulted in a low

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.

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.

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

Table 3: 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 AI. For each evaluated attribute, we report its mean Likert scale and the 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 ₆₂	−0.03 ₄₈	0.03 ₆₁	0.02 ₅₄	−0.03 ₅₁	−0.02 ₆₀
Informed Human	0.01 ₆₄	0.02 ₅₃	0.03 ₅₈	0.06 ₅₅	−0.02 ₅₇	0.10 ₆₁

Table 4: We report the inter-annotator agreement (IAA) among five annotators across **six different metrics** using Krippendorff’s α . The subscript in the IAA column (%) is used to denote the average percentage of annotators who agree on the most voted rating.

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.

inter-annotator agreement (IAA) in Krippendorff’s α among the five annotators. We have included the IAA scores for six distinct metrics in Table 4. Additionally, the IAA scores for five personality traits are presented in Table 5.

C LLM as Evaluators

C.1 Temperature

We experiment with different temperatures with the ChatGPT and GPT-4 evaluators and observe similar trends reported by Chiang and Lee (2023). As shown in Table 6, we see the ratings given by

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.

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
○ ○ ○ ○ ○

Figure 15: Agreeableness question on Prolific.

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

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

Table 5: We report the inter-annotator agreement (IAA) among five annotators across **five personality traits** using Krippendorff’s α . The subscript in the IAA column (%) is used to denote the average percentage of annotators who agree on the most voted rating.

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

Table 6: 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.

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)



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)



Figure 17: Openness question on Prolific.

for future research.

D BFI Scores and Personality Traits

In addition to report the significant LIWC features correlated with the binary label in the main paper, we also conduct similar study with the orig-

inal 5 point labels with Spearsman’s ρ and report our findings here.

D.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 shows 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 shows 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 analytic 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”.

D.2 GPT-4 Personas

Extroversion Introverted personas have more descriptions of the 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 shows 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 has

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 involves 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 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 that have 7 letters or longer and more words per sentence (BigWord: $\rho = 0.29, p < .001$; WPS: $\rho = 0.26, p < .001$).