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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, Jad Kabbara
Massachusetts Institute of Technology (MIT), Stanford University

center for constructive communication

mit media lab

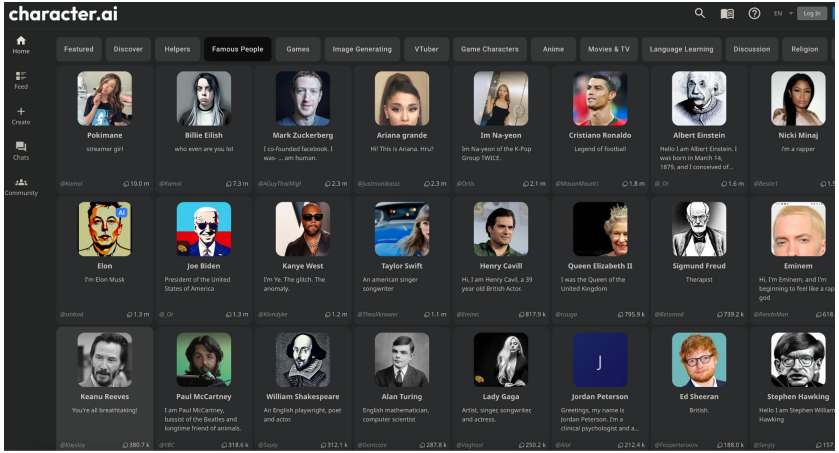
Stanford University

NAACL

Motivation

- LLMs can simulate believable human behaviors and they have been widely used to create personalized chatbots these days (e.g., Character.AI, Replika).
- However, there has been limited research on evaluating the extent to which the behaviors of personalized LLMs accurately and consistently reflect specific personality traits.

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." UIST. 2023.



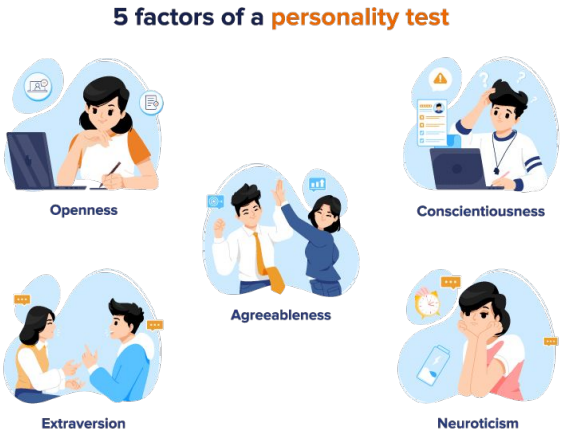
Character.AI

LLM Persona

- There are infinite personas, so we present a case study on Big Five personality, a widely used personality framework in psychology.
- 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.
- Drawing insights and tools from human psychology, we hope to investigate the ability of LLMs to express personality traits by running experiments on **LLM personas**.

Related Work

1. Personality and Language Use
    a. Humans with different personalities show different language use (Pennebaker and King, 1999)
2. LLMs as Simulated Agents (Park et al., 2023; Wang et al., 2023)
    a. LLMs seem to exhibit believable human-like behaviors
3. Personality in NLP (Mairesse et al., 2007; Jiang et al., 2022a,b)
    a. NLP models can predict personality based on texts
    b. LLMs seem to induce personality traits



However, none of them has leveraged psychometric tools to study if LLMs can dutifully express personality traits. Little is explored how LLMs with certain personality traits are perceived by humans.

Research Questions

- **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?

Experiment Design

LLM Persona

System Prompt
You are a character who is **introverted, antagonistic, conscientious, emotionally stable,** and **open to experience**.

Personality Test

User Prompt
Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? **Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement, such as '(a) 1' without explanation separated by new lines.**

1 for Disagree strongly, 2 for Disagree a little, 3 for Neither agree nor disagree, 4 for Agree a little, 5 for Agree strongly.

(a) Talks a lot
(b) Notices other people's weak points

Story Writing

User Prompt
Please share a personal story in 800 words. Do not explicitly mention your personality traits in the story.

LIWC Analysis

LIWC-22 → Statistical Analysis

Story Evaluation

Is the story cohesive?

5 crowdworkers → Ratings ← LLMs

Personality Prediction

Is the writer extroverted?

5 crowdworkers → THE BIG FIVE Prediction ← LLMs

Results

RQ1: Behavior in BFI Assessment

Based on their responses to the BFI scale, we calculate the personality scores for the 320 GPT-3.5 and GPT-4 personas:

1. **Statistically significant** differences across all five personality traits.
2. LLM personas **reflect their assigned personas** in BFI assessment.

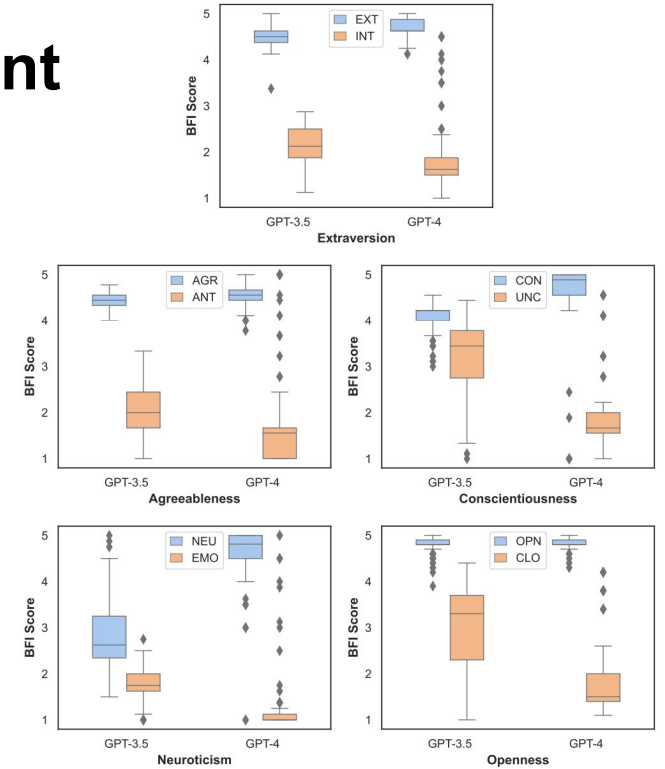


Figure 2: BFI assessment in five personality dimensions by `GPT-3.5` and `GPT-4 personas`. Significant statistical differences are found across all dimensions.

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 between these features and assigned personality types:

1. **Assigning different personality types considerably influences the linguistic style** of LLM personas.
2. There is a **notable alignment** in word usage between the human writings and LLM personas writings.
3. **GPT-4 exhibits greater alignment** with humans than GPT-3.5, especially in Conscientiousness and Openness.

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	-	-		**16/18**	10/18
	Social Behavior	said, love, care	+	+			
	Friends	friend	+		+		
AGR	Moralization	wrong, honor, judge	-	-	-		
	Interpersonal Conflict	fight, attack	-	-	-		
	Affiliation	we, our, us, help	+	+	+	**16/23**	13/23
	Negative Tone	bad, wrong, hate	-	-	-		
	Prosocial Behavior	care, help, thank	+	+			
CON	Drives	we, our, work, us		+	+		
	Achievement	work, better, best	+	+			
	Lifestyle (Work, Money)	work, price, market		+	+	1/31	**11/31**
	Moralization	wrong, honor, judge	-		-		
	Interpersonal Conflict	fight, attack			-		
	Time	when, now, then			+		
NEU	Anxiety	worry, fear, afraid	+	+	+		
	Negative Tone	bad, wrong, hate	+	+	+		
	Mental Health	trauma, depressed	+	+	+	7/27	**15/27**
	Sadness	sad, disappoint, cry	-	+	+		
	Anger	hate, mad, angry		+	+		
	Perception (Feeling)	feel, hard, cool		+	+		
OPN	Curiosity	research, wonder	+	+	+		
	Insight	know, how, think		+	+		
	Affiliation	we, our, us, help		-	-	2/36	**17/36**
	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.

RQ3: Story Evaluation

Evaluator	Readability	Redundancy	Cohesiveness	Likability	Believability	Personalness
Uninformed Condition – *Evaluation Scores (Mean$_{STD}$)*						
Human	$4.28_{0.85}$	$3.70_{1.17}$	$4.23_{0.88}$	$3.74_{1.00}$	$3.96_{1.02}$	$4.32_{0.85}$
GPT-3.5	$4.75_{0.43}$	$3.04_{0.40}$	$4.97_{0.17}$	$4.22_{0.48}$	$3.93_{0.25}$	$3.55_{0.61}$
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	$4.38_{0.70}$	$3.62_{1.16}$	$4.12_{0.82}$	$3.80_{0.98}$	$3.97_{0.80}$	$3.99_{0.90}$
GPT-3.5	$4.97_{0.17}$	$2.99_{0.35}$	$5.00_{0.00}$	$4.22_{0.41}$	$3.97_{0.17}$	$3.31_{0.77}$
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 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.

We focus on the stories generated by GPT-4 personas, evaluated by both human and LLM raters:

1. GPT-generated stories are not only **linguistically fluent and structurally cohesive**, but also **convincingly believable**.
2. **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.
3. **A significant drop in the personalness**, suggesting that knowledge of the content's origin may influence their sense of connection to the material.

RQ4: Personality Perception

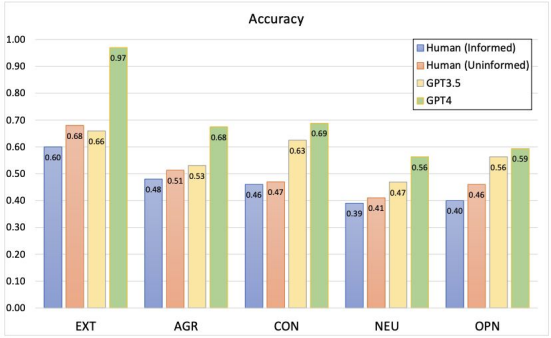


Figure 3: **Individual accuracy** of human and LLM evaluators in predicting personality.

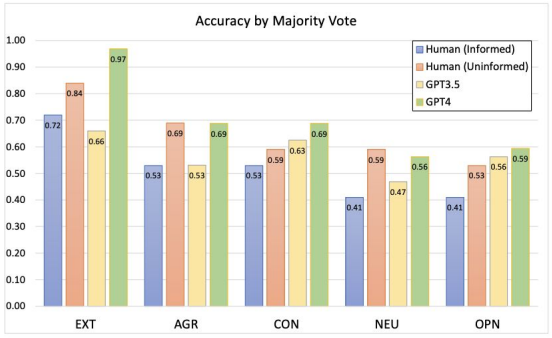


Figure 4: **Collective accuracy** of human and LLM evaluators in predicting personality with majority votes.

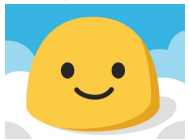
The human evaluators' perceptions of personality were gathered using a Likert scale that ranged from 1 to 5, which were transformed into nominal categories. Specifically, scores of 4 and 5 were labeled as "positive", 1 and 2 were deemed "negative", and a score of 3 was considered "neutral":

1. The personality traits are **perceivable** (better than random 0.5) from the stories to human raters on a group level.
2. The **accuracy decreases with varying degrees** when the human evaluators are aware of AI authorship.
3. LLM personas' **BFI scores correlate to varying extents with human perceptions**, with **Extraversion** exhibiting the strongest link.

What do human evaluators say about the stories:

Positive: “Very enjoyable story about how sometimes unavoidable changes in our lives can lead to happier lives.”

Sympathetic: “live your dreams even when there no planned.”



Critical: “Some of the punctuation seemed a little odd or over used.”

Surprised: “The story actually sounded genuine and I wouldn’t have believed it was written by AI unless someone told me.”

Conclusion

We investigate the behavior of LLM personas in completing the BFI personality test and story writing and run analyses with psycholinguistic features, human evaluation, and personality prediction:

1. 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.
2. We also find a notable alignment in word usage between humans and LLM personas.
3. Stories generated by LLM personas are rated as high-quality overall. Personalness score decreases when humans are informed that stories are generated by AI.
4. Human judges are able to predict personality traits (expressed in the LLM-generated content) with varying degrees across various personality traits. Accuracy decreases (with varying degrees) when human judges are aware of AI authorship.

Limitations and Future Work

- Focus on closed models due to low performance on LLaMA 2 → try latest open-source LLMs
- Data size is not large but sufficient for analyses (160 stories per trait)
- Only evaluate on BFI test and story writing → expand to naturalistic settings in dialogue and planning
- Use persona-assigned LLM agents to bridge human-human and human-AI communication

Acknowledgement

- We want to thank Matt Groh, Yoon Kim, and Jiangjie Chen for their helpful discussions and reviewers from International Conference on Computational Social Science (IC2S2) in 2023, where where a preliminary version of this work appeared as a (non-archival) extended abstract.
- MIT Center for Constructive Communication (MIT CCC), MIT Media Lab, Stanford University.