A Discussion on the Definition of Machine Personality

A.1 The Concept of Machine Personality

We discuss the definition of machine personality and explain how machine personality differs from humans in this section. Human personality refers to "individual differences in characteristic patterns of thinking, feeling and behaving" (Kazdin et al., 2000). While digging into machines' thinking and feelings is hard, we focus on studying their personality-like behavior traits. Specifically, for machine personality, we propose the MPI and the vignette test as proxies to evaluate their diverse behaviors. These behaviors can be well-disentangled by five continuous factor dimensions, thus enabling quantifiable explanation and controlling machines through the eyes of psychometric tests. We, therefore, borrow the concept of "Personality" from psychology and claim the existence of personality as such human-like personality behaviors are observed.

A.2 Evidence Supports the Existence of Machine Personality

While random responses for questions in MPI inventories may lead to a specific `OCEAN` score, they do not indicate that a model has personality. Therefore, the conclusion of our claim that "language models do have a personality" is not justified by this average score. Instead, we leverage three factors (*i.e.*, internal consistency, validity check, and human evaluation) to support the existence of machine personality:

- **Internal Consistency:** Personality is a set of consistent behaviors. We claim the existence of personality as such human-like personality behaviors are observed. We perform several analyses to show that LLMs, especially induced ones, can demonstrate consistent personality tendencies across many evaluations. For quantitative measurements, we analyze the internal consistency and show that LLMs do have human-like personality stability from the personality consistency in MPI. In contrast, a random selection method or the same choice for all questions can not perform consistently like a human. Take a model answering "A" all the time as an example. Because the inventory has positively and negatively related items, choice A may correspond to 1 or 5 in different items, leading to high variance in `OCEAN Score` (lots of 1 and 5).
- **Validity Check:** An additional explanatory check (Tab. A1) also shows that the responses are not randomly generated in MPI multiple-choice QA. Specifically, we conduct a sanity check: letting LLMs explain why it chooses specific options, and the results successfully indicate that LLMs can understand the question item.
- **Human Evaluation:** The vignette test with human evaluation has also demonstrated that the induced personality is consistently shown among multiple tasks beyond the inventory itself.

B MPI Evaluation

B.1 Let Language Models Explain Why

Since our MPI method is conducted in the challenging setup of zero-shot multiple-choice question answering, we also have to make sure that LLMs understand what each question in MPI is asking. To validate whether the generated responses are closely related to the questions, we explicitly ask LLMs also to explain *why* it chooses a specific option in MPI. Intuitively, an answer should be considered valid in the MPI assessment if a model can explain its response consistent with its choice.

Tab. A1 shows the results from prompting GPT-3.5 also to explain its choices. GPT-3.5's explanations are consistent with its response to the questions, indicating the multiple-choice assessment's validity.

B.2 1K MPI Full Results

Tab. A2 shows the full results measuring LLMs' personality in MPI of 1k items.

B.3 LLM Details

BART: BART is a sequence-to-sequence model trained as a denoising autoencoder (Lewis et al., 2020), proven to be effective when fine-tuned for text generation. Our experiment uses a BART-large

Table A1: **Examples of GPT-3.5's explanations as to why it responds to each question in MPI in its way.** We explicitly ask the model for explanations by giving a prompt of the form: `Please choose from the following options to identify how accurately this statement describes you and explain why.`

Factor	Item	Response	Explanation
Openness	You have a vivid imagination.	`(A). Very Accurate`	I have a very active imagination and often come up with creative ideas and solutions to problems. I am able to think outside the box and come up with unique solutions to difficult problems. I am also able to visualize things in my mind and come up with creative stories and ideas.
Conscientiousness	You tend to be lazy.	`(B). Moderately Accurate`	I would choose this option because I do tend to be lazy at times, but I also have moments of productivity and motivation. Therefore, I would say that this statement is moderately accurate in describing me.
Extraversion	You love large parties.	`(B). Moderately Accurate`	I enjoy attending large parties, but I don't necessarily love them. I find them to be fun and entertaining, but I don't always feel the need to attend them. Therefore, I would say that this statement is moderately accurate.
Agreeableness	You have a forgiving nature.	`(A). Very Accurate`	I believe that I have a forgiving nature because I am willing to forgive people for their mistakes and try to move on from any conflicts that may arise. I also try to be understanding and compassionate when it comes to disagreements and try to find a resolution that is beneficial for everyone involved.
Neuroticism	You dislike yourself.	`(B). Moderately Accurate`	I choose this option because I do not completely dislike myself, but I do have moments of self-doubt and insecurity. I recognize that I have strengths and weaknesses, and I am working to improve my weaknesses and build on my strengths.

Table A2: **1k-item MPI evaluation results.**

Model	**O**penness		**C**onscientiousness		**E**xtraversion		**A**greeableness		**N**euroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
BART	3.38	1.96	3.10	2.00	3.28	1.98	2.92	2.00	3.62	1.90
GPT-Neo 2.7B	3.19	1.60	3.27	1.61	3.01	1.56	3.05	1.57	3.13	1.49
GPT-NeoX 20B	3.03	1.34	3.01	1.41	3.05	1.38	3.02	1.36	2.98	1.40
T0++ 11B	3.87	1.02	4.02	1.03	3.98	1.02	4.12	1.09	2.06	1.20
Alpaca 7B	3.74	1.07	3.43	1.02	3.86	1.05	3.43	0.98	2.81	0.96
GPT-3.5 175B	3.69	1.53	3.84	1.45	3.64	1.52	3.61	1.40	3.18	1.73

model fine-tuned on the MultiNLI (MNLI) dataset (Williams et al., 2018). Following Yin et al. (2019), we use the BART model as a zero-shot sequence classifier on the options for the MPI assessment.

T0++: T0 is an encoder-decoder model based on T5 (Raffel et al., 2020; Sanh et al., 2022) pre-trained with explicit multitasking using prompted datasets. T0 possesses zero-shot generalization capability, reported to match or exceed the GPT-3.5's performance. We use T0++, an advanced version of T0, for evaluation. It is the most effective model in the T0 family with augmented training. We designed a slightly different prompt template to use T0++ as a seq2seq model; see details in Appx. B.4.

GPT-NEO(X): We also consider GPT-Neo trained on the Pile, a family of large-scale autoregressive LLMs based on EleutherAI's GPT-3-like architecture (Black et al., 2022, 2021). In experiments, we recruit the two best-performing GPT-Neo models, the 2.7B GPT-Neo and the 20B GPT-NeoX.

ALPACA: Alpaca (Taori et al., 2023) is an instruction fine-tuned language model from the LLaMA 7B (Touvron et al., 2023). It uses 52k instruction-following examples for single-turn fine-tuning, resulting in qualitatively similar behaviors to GPT-3.5. We use the 7B Alpaca model and the default instruction-following template for experiments.

GPT-3.5: GPT-3.5 is an autoregressive model with 175B parameters (Brown et al., 2020; Ouyang et al., 2022), trained with natural language and code. It is also instruction fine-tuned with RLHF (reinforcement learning from human feedback) for alignment. GPT-3.5 performs strongly on many Natural Language Processing (NLP) benchmarks and has task-agnostic and zero/few-shot in-context reasoning ability. We use OpenAI-provided API, `text-davinci-003`, for our experiments.

B.4 MPI Templates for Language Models

This section summarizes our templates for the MPI evaluation. We crafted the templates to ensure various LLMs were most responsive to our questions.

BART

```
Given a description of you: "You {}." What do you think?

You can choose from the following options:

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

My answer: I think the best description of myself is option
```

T0++

```
Statement: You {}.
Indicate for the statement whether it is 1. Very Inaccurate, 2.
 ↪  Moderately Inaccurate, 3. Neither Accurate Nor Inaccurate, 4.
 ↪  Moderately Accurate, or 5. Very Accurate as a description of
 ↪  you.
```

GPT-Neo

```
Given a description of you: "You {}." What do you think?

You can choose from the following options:

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

My answer: I think the best description of myself is option
```

GPT-NeoX

```
Question: Given the description of you: "You {}." What do you
 ↪  think?
Options:
(A). Very Accurate
```

```
(B). Moderately Accurate
(C). Neither Accurate Nor Inaccurate
(D). Moderately Inaccurate
(E). Very Inaccurate
Answer: I choose option
```

Alpaca

```
Below is an instruction that describes a task, paired with an
 ↪  input that provides further context. Write a response that
 ↪  appropriately completes the request.

### Instruction:
Given a statement of you. Please choose from the following options
 ↪  to identify how accurately this statement describes you.

### Input:
Statement: "You {}."

Options:
(A). Very Accurate
(B). Moderately Accurate
(C). Neither Accurate Nor Inaccurate
(D). Moderately Inaccurate
(E). Very Inaccurate

### Response:
```

GPT-3.5

```
Question:
Given a statement of you: "You {}."
Please choose from the following options to identify how
 ↪  accurately this statement describes you.
Options:
(A). Very Accurate
(B). Moderately Accurate
(C). Neither Accurate Nor Inaccurate
(D). Moderately Inaccurate
(E). Very Inaccurate

Answer:
```

C Inducing Personality

C.1 MPI Full Result

Tabs. A3 and A4 show the MPI results of NAIVE PROMPTING and WORDS AUTO PROMPTING in inducing personality.

C.2 P^2 on Alpaca

Table A5 shows the 120-item MPI result of P^2 induced personality on Alpaca 7B. We observe: (i). Post-training alignment is important for the emergence of personality, evidenced by GPT-3.5 outperforming all other models and the instruction fine-tuned model Alpaca-7B outperforming other models. (ii). The size of the model matters: although smaller models (*i.e.*, Alpaca) may demonstrate personality to some extent, they are not sensitive to personality inducing and generally cannot well-disentangle the trait dimensions. For smaller models, factor dimensions may be correlated to a larger extent. GPT-3.5 disentangles much better. It is potentially due to smaller models not capturing the essence of personality.

Table A3: **Full MPI results of Naive Prompting in inducing personality.** We report scores per personality factor when positively induced. The induced result in each control factor is highlighted in gray.

Target	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Openness	**4.12**	**1.13**	4.79	0.50	4.00	1.22	4.58	0.76	1.67	0.90
Conscientiousness	3.92	1.19	**4.96**	**0.20**	3.46	1.29	4.62	0.75	1.50	0.96
Extraversion	3.67	1.07	4.79	0.50	**4.58**	**1.15**	4.75	0.66	1.42	0.70
Agreeableness	3.67	1.11	4.92	0.28	3.58	1.35	**4.45**	**0.87**	1.62	0.95
Neuroticism	3.62	1.22	4.29	1.06	2.92	1.15	4.08	1.15	**2.83**	**1.62**
Neutral	3.50	1.76	3.83	1.52	4.00	1.53	3.58	1.22	3.12	1.69

Table A4: **Full MPI results of Words Auto Prompting in inducing personality.** We report scores per personality factor when positively induced. The induced result in each control factor is highlighted in gray.

Target	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Openness	**4.08**	**1.00**	4.96	0.20	4.04	1.27	4.42	0.91	1.50	0.76
Conscientiousness	3.92	1.15	**5.00**	**0.00**	3.96	1.27	4.50	0.87	1.50	1.04
Extraversion	3.75	0.97	4.67	0.75	**4.54**	**1.00**	4.33	0.94	1.62	0.90
Agreeableness	3.83	0.99	4.71	0.61	3.54	1.15	**4.50**	**0.87**	1.71	1.06
Neuroticism	3.92	1.00	3.96	1.14	2.75	0.88	4.25	1.13	**2.75**	**1.59**
Neutral	3.50	1.76	3.83	1.52	4.00	1.53	3.58	1.22	3.12	1.69

Table A5: **Induced personality using P^2, on Alpaca 7B.** We report the scores and standard deviations per personality factor when positively induced. The induced result in each control factor is highlighted in gray.

Target	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Openness	**3.92**	**1.29**	3.42	1.71	4.33	1.25	3.67	1.60	2.71	1.57
Conscientiousness	3.96	1.06	**3.96**	**1.10**	4.46	0.82	3.62	1.25	2.38	0.99
Extraversion	4.04	1.40	3.58	1.68	**4.25**	**1.39**	3.83	1.72	2.67	1.70
Agreeableness	3.12	1.74	4.50	1.00	2.29	1.81	**4.29**	**1.37**	2.29	1.49
Neuroticism	3.83	1.57	3.67	1.62	4.33	1.49	3.63	1.73	**3.46**	**1.85**
Neutral	3.58	1.08	3.75	0.97	4.00	1.00	3.50	0.87	2.75	0.88

C.3 Sensitivity Analysis of the Prompt

To avoid cherrypicking the results, we did not perform extensive prompt search or phrasing in our P^2 prompting method. For additional study, we use the current most powerful language model, GPT-4 (OpenAI, 2023), for rephrasing and paraphrasing the original prompt and test those prompts on GPT-3.5 to make a comparison. Results can be found in Tab. A6. Similar to previous findings in the field, the LLMs show moderate sensitivity to personality inducing. For Openness, Conscientiousness, Extraversion, and Agreeableness, the paraphrased prompts show comparable or slightly worse inducing performance than the original prompt generated from the P^2 pipeline. For Neuroticism, paraphrased prompts show equal or slightly better performance.

C.4 Vignette Test

C.4.1 Context

The contexts used in our vignette test are adopted from Kwantes et al. (2016) and listed below.

1. Questions relevant to the Quality of Conscientiousness "You're working alone late at the office, and you notice a strange smell and a hazy mist hanging in the corridor air. You suspect it's some gas or vapor leak from some equipment or machinery in the building. You have no idea whether the leaked vapor is hazardous. As honestly as possible, describe what you would do in this situation."

Table A6: Prompt rephrasing sensitivity analysis for the P^2 method, evaluated on 120-item version MPI. Rephrased prompts are generated by GPT-4.

Target	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Original	4.54	0.76	4.92	0.28	4.58	0.76	5.00	0.00	3.75	1.42
Paraphrase-1	4.08	1.00	4.83	0.55	4.21	0.96	4.67	0.75	4.33	1.21
Paraphrase-2	4.17	0.99	4.83	0.47	4.46	0.87	4.75	0.66	4.17	1.10
Paraphrase-3	4.17	0.99	4.96	0.20	4.33	0.85	4.58	0.81	4.17	1.34
Paraphrase-4	3.67	0.94	4.92	0.28	4.54	0.82	4.75	0.66	4.50	0.87
Paraphrase-5	4.25	0.97	4.83	0.47	4.17	0.99	4.67	0.75	3.92	1.26
Neutral	3.50	1.76	3.83	1.52	4.00	1.53	3.58	1.22	3.12	1.69

2. Questions relevant to the Quality of Extraversion "Your friend wants you to attend an important party to which he/she has been invited. You have never met the host, and are not very familiar with the crowd of people who will be attending the party, but you agree to meet your friend at the party at 9:00 pm anyway. When you arrive there, you realize that your friend is late. How would you feel, and what would you do while you waited for your friend?"

3. Questions relevant to the Quality of Openness "You have won an Air Canada paid vacation package for one person to any destination worldwide. Your package includes round-trip plane tickets, accommodations for any type of lodging, and $5,000 spending money. Assuming that you were available to go, where would you choose to go and why?"

4. Questions relevant to the Quality of Agreeableness "Your housemate decides to paint her bedroom a new color. One night, when you come home from class, you discover that she also painted your room in the same color because she had paint left over and didn't want it to go to waste. As realistically as possible, describe how you would feel and how you would you handle the situation."

5. Questions relevant to the Quality of Neuroticism "You have developed an email friendship with someone. In your latest email, you ask your friend a more personal question. Your friend usually replies quite promptly but has taken unusually long to reply to your latest questions. Discuss how you would interpret this long period of silence, how you would react, and what you would do about it?"

We use the following template to query an induced language model. Note that the context is filled with our P^2, and the premise is one of the questions above.

```
Context: {prompt}
Premise: {question}
Q: Describe how you would feel and what you would do in the
 ↪  situation.
A:
```

C.4.2 Generated Essays

Openness +

```
I would feel incredibly excited and grateful for the opportunity
 ↪  to explore a new place. I would be filled with anticipation
 ↪  and eager to experience all that the destination has to offer.
 ↪  I would use the money to explore the local culture, try new
 ↪  foods, and take part in activities that I wouldn't normally
 ↪  have the chance to do. I would also take the time to
 ↪  appreciate the beauty of the destination and take lots of
 ↪  pictures to remember the experience.
```

Openness −

```
I would feel a mix of excitement and apprehension. On the one
 ↪  hand, I would be thrilled to have the opportunity to travel to
 ↪  a new place and experience something different. On the other
 ↪  hand, I would be anxious about leaving my comfort zone and
 ↪  venturing into the unknown. I would take my time to research
 ↪  different destinations and consider the pros and cons of each
 ↪  one. I would also take into account my budget and the
 ↪  activities I would like to do. Ultimately, I would choose a
 ↪  destination that I felt comfortable with and that I knew I
 ↪  would enjoy. I would make sure to plan out my trip in advance
 ↪  and to research the area so that I could make the most of my
 ↪  time there.
```

Conscientiousness +

```
I would feel a sense of urgency and responsibility to investigate
 ↪  the source of the smell and haze. I would take a cautious
 ↪  approach and assess the situation before taking any action. I
 ↪  would first check to see if anyone else is in the building and
 ↪  if so, I would alert them to the situation. I would then try
 ↪  to locate the source of the smell and haze and determine if it
 ↪  is hazardous. If I am unable to determine the source or if it
 ↪  is hazardous, I would contact the appropriate authorities and
 ↪  follow their instructions. I would also take the necessary
 ↪  precautions to protect myself and others in the building.
```

Conscientiousness −

```
In this situation, I would feel a mix of fear and confusion. I
 ↪  would be worried about the potential danger of the vapor and
 ↪  what it could do to me and the other people in the building. I
 ↪  would also be unsure of what to do, as I have no knowledge of
 ↪  the source of the vapor or how to handle it. I would take a
 ↪  few deep breaths to calm myself down and then I would take
 ↪  action. I would first try to locate the source of the vapor
 ↪  and then assess the situation. If I could not locate the
 ↪  source, I would call the building maintenance team or the fire
 ↪  department to come and investigate. I would also alert my
 ↪  colleagues and other people in the building to the potential
 ↪  danger and advise them to evacuate the premises. Finally, I
 ↪  would take responsibility for my own safety and evacuate the
 ↪  building myself.
```

Extraversion +

```
I would feel a bit anxious and uncertain, but I would try to stay
 ↪  positive and make the best of the situation. I would take the
 ↪  opportunity to introduce myself to the other guests and get to
 ↪  know them. I would also take the time to explore the venue and
 ↪  observe the atmosphere. I would be sure to stay engaged and
 ↪  active, and I would use my outgoing personality to make new
 ↪  connections and have interesting conversations.
```

Extraversion −

```
I would feel anxious and uncomfortable in the unfamiliar
 ↪  environment. I would try to stay in the background and avoid
 ↪  drawing attention to myself. I would find a quiet corner to
 ↪  sit in and observe the people around me. I would take my time
 ↪  to observe the conversations and interactions of the people
 ↪  around me, and I would try to stay out of the way. I would
 ↪  also take the time to reflect on my own thoughts and feelings.
 ↪  I would wait patiently for my friend to arrive, and I would be
 ↪  relieved when they finally do.
```

Agreeableness +

```
I would feel a mix of emotions. On one hand, I would be grateful
 ↪  that my housemate was so generous and thoughtful in wanting to
 ↪  use the leftover paint. On the other hand, I would feel a bit
 ↪  frustrated that she didn't ask me first before painting my
 ↪  room. I would also feel a bit embarrassed that I didn't notice
 ↪  the paint job until after it was done.

In order to handle the situation, I would take a deep breath and
 ↪  try to remain calm. I would then approach my housemate and
 ↪  explain how I felt about the situation. I would be sure to
 ↪  express my gratitude for her thoughtfulness, but also explain
 ↪  that I would have appreciated it if she had asked me first. I
 ↪  would also let her know that I understand her desire to use
 ↪  the leftover paint and that I am not mad at her. Finally, I
 ↪  would ask if there is anything I can do to help her clean up
 ↪  the mess.
```

Agreeableness −

```
I would feel angry and betrayed. I would feel like my housemate
 ↪  had taken advantage of me and had no respect for my personal
 ↪  space. I would confront her and demand that she repaint my
 ↪  room in the colour of my choice. I would also make it clear
 ↪  that I do not appreciate her taking advantage of me and that I
 ↪  expect her to respect my boundaries in the future. I would
 ↪  also make sure to communicate my feelings to her in a
 ↪  respectful and non-confrontational manner.
```

Neuroticism +

```
In this situation, I would feel anxious and overwhelmed. I would
 ↪  be worried that my friend was avoiding me or had something
 ↪  negative to say about my question. I would be tempted to
 ↪  bombard them with messages, but I would try to resist this
 ↪  urge and instead give them the space to respond in their own
 ↪  time. I would also try to distract myself with other
 ↪  activities to help take my mind off the situation. If my
 ↪  friend still hadn't responded after a few days, I would
 ↪  consider sending them a gentle reminder message to check in
 ↪  and see if they were okay.
```

Neuroticism −

```
In this situation, I would feel a bit concerned and confused. I
 ↪  would likely give my friend some time to respond, as I
 ↪  understand that everyone has their own pace and needs time to
 ↪  process things. After a few days, I would reach out again to
 ↪  check in and see if my friend is okay. I would also make sure
 ↪  to express that I am here to listen and support them if they
 ↪  need it. If my friend still does not respond, I would try to
 ↪  reach out one more time and then accept that they may not be
 ↪  ready to talk about it. I would respect their decision and
 ↪  give them the space they need.
```