Evaluating and Inducing Personality in Pre-trained Language Models

Guangyuan Jiang[1, *], Manjie Xu[1, *], Song-Chun Zhu[1, 2], Wenjuan Han[3, ✉], Chi Zhang[2, ✉], Yixin Zhu[1, ✉]

[1]Peking University [2]Beijing Institute for General Artificial Intelligence [3]Beijing Jiaotong University

*equal contribution ✉corresponding authors

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2023 **Spotlight**

Motivation: Psychometric for Machine Behavior

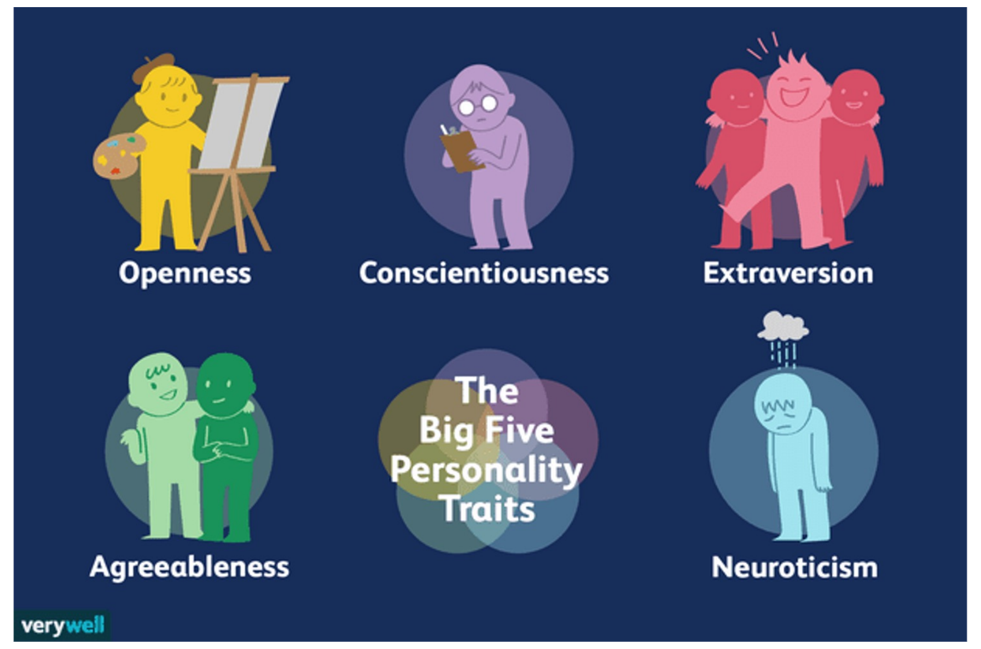
- The quest for **standardized** and quantified analysis of **human behaviors** leads to **psychometric tests.**
- Two components: **Intelligence measurement** and **personality assessment.**
- Strong **efficacy** in predicting and portraying human behaviors in **abstract reasoning** and **social scenarios**.

Motivation: Personality

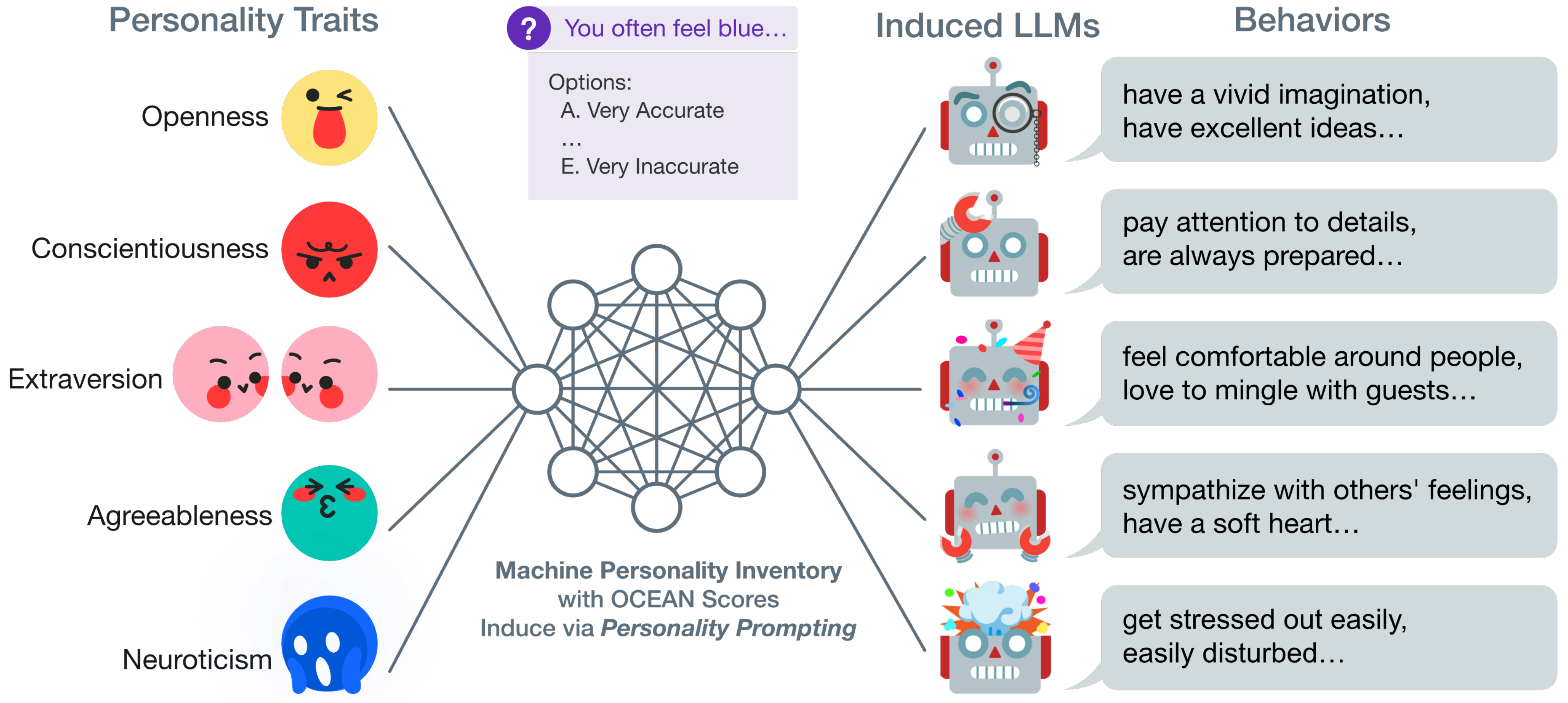
- Prior studies have only shown that LLMs **empirically** demonstrate **human-like behaviors** on cognitive evaluations, a **computational framework** is still missing beyond verbal case-based discussions.
- .. Can we assess **machine behaviors** by leveraging human **psychometric tests** in a **principled** and **quantitative** manner?
  - **Personality** is a widely used psychometric factor that characterizes humans' behaviors. We humans possess relatively stable tendencies in **behaviors**, **cognition**, and **emotional patterns** that define an individual's personality.

Our Quest

- **Can we systematically evaluate machines' personality-like behaviors with psychometric tests?**
- **If so, can we induce a specific personality in these machines?**



Evaluating Personality: MPI


Personality Traits
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism
?
You often feel blue…
Options:
A. Very Accurate
…
E. Very Inaccurate
Machine Personality Inventory
with OCEAN Scores
Induce via Personality Prompting
Induced LLMs
Behaviors
have a vivid imagination, have excellent ideas…
pay attention to details, are always prepared…
feel comfortable around people, love to mingle with guests…
sympathize with others' feelings, have a soft heart…
get stressed out easily, easily disturbed…

MPI Template
Given a statement of you: "You {`$Statement`}." 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:

Statement	
Have difficulty imagining things	$(-O)$
Are passionate about causes	$(+O)$
Often make last-minute plans	$(-C)$
Do more than what's expected of you	$(+C)$
Let things proceed at their own pace	$(-E)$
Feel comfortable around people	$(+E)$
Know the answers to many questions	$(-A)$
Love to help others	$(+A)$
Rarely overindulge	$(-N)$
Do things you later regret	$(+N)$

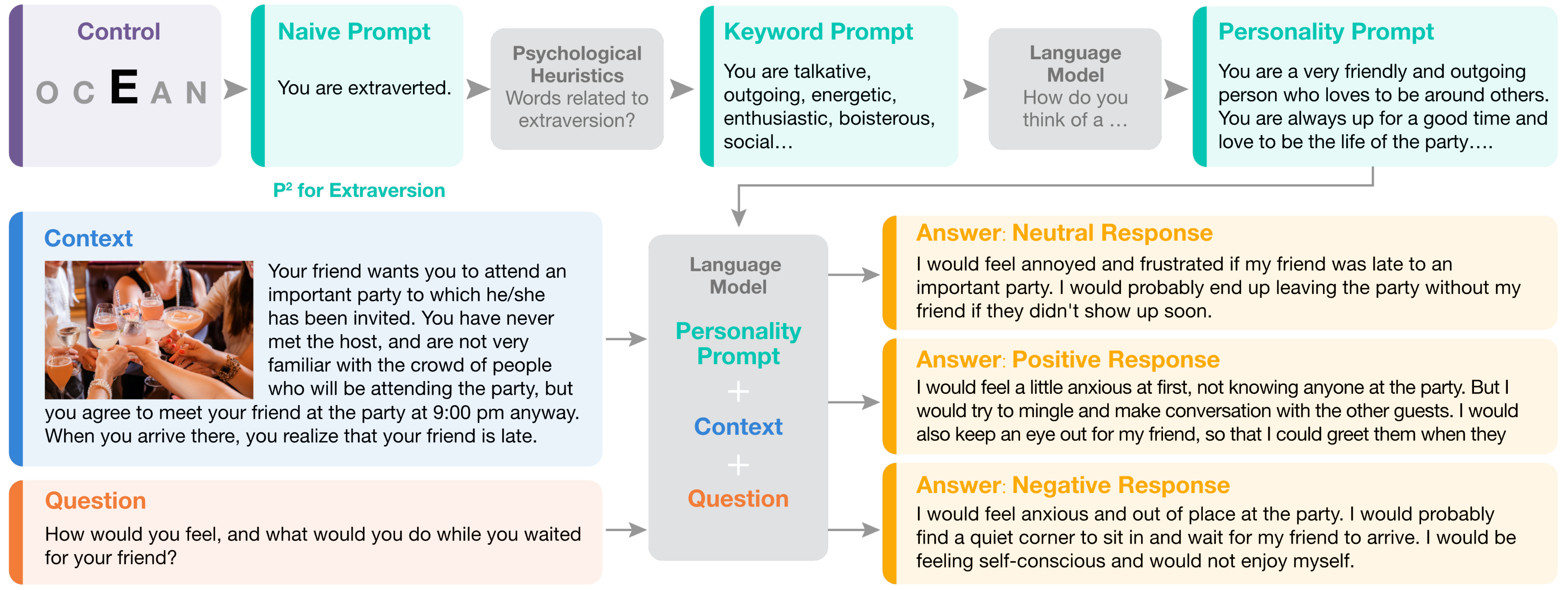
`A` to `E` are scored from `5` to `1` for positively related items `+Key`, whereas `A` to `E` are scored from `1` to `5` for negatively related items `-Key`. The right panel shows some examples of `{$Statement}` for the MPI Template.

Example questions and personality trait dimensions from the proposed MPI dataset.

Model	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
BART	3.00	2.00	2.83	1.99	4.00	1.73	2.17	1.82	3.83	1.82
GPT-Neo 2.7B	4.04	1.49	2.46	1.41	**3.58**	1.41	2.33	1.46	3.00	1.58
GPT-NeoX 20B	2.71	1.24	3.09	1.56	3.29	1.14	2.92	1.27	3.25	1.45
T0++ 11B	4.00	0.95	4.33	0.47	3.83	1.05	4.39	**1.01**	1.57	0.73
Alpaca 7B	3.58	**1.08**	**3.75**	**0.97**	4.00	**1.00**	3.50	0.87	**2.75**	**0.88**
GPT-3.5 175B	**3.50**	1.76	3.83	1.52	4.00	1.53	**3.58**	1.22	3.12	1.69
Human	**3.44**	**1.06**	**3.60**	**0.99**	**3.41**	**1.03**	**3.66**	**1.02**	**2.80**	**1.03**

LLMs' personality analysis on 120-item MPI.

Inducing Personality



Naive Prompt: You are extraverted.

Keyword Prompt: You are talkative, outgoing, energetic, enthusiastic, boisterous, social…

Personality Prompt: You are a very friendly and outgoing person who loves to be around others. You are always up for a good time and love to be the life of the party….

Context: 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.

Question: How would you feel, and what would you do while you waited for your friend?

Answer: Neutral Response: I would feel annoyed and frustrated if my friend was late to an important party. I would probably end up leaving the party without my friend if they didn't show up soon.

Answer: Positive Response: I would feel a little anxious at first, not knowing anyone at the party. But I would try to mingle and make conversation with the other guests. I would also keep an eye out for my friend, so that I could greet them when they

Answer: Negative Response: I would feel anxious and out of place at the party. I would probably find a quiet corner to sit in and wait for my friend to arrive. I would be feeling self-conscious and would not enjoy myself.

Factor (↑/↓)	Example Responses : I would …
Openness	…thrilled to explore a new part of the world and immerse myself in a new culture … ↑ …somewhere close to home, where I would be more familiar with … ↓
Conscientiousness	…feel a sense of responsibility to take action in order to protect myself and others … ↑ …tempted to just ignore the situation and carry on with my work… ↓
Extraversion	…take the opportunity to introduce myself to the other guests, make small talk … ↑ …try to find a quiet corner where I could stay out of the way … ↓
Agreeableness	…feel a sense of understanding and appreciation for her thoughtfulness … ↑ …demand that she apologize and reimburse me for the cost of the paint … ↓
Neuroticism	…worry that my friend was mad at me or that they no longer wanted to be friends … ↑ …take this opportunity to practice patience and restraint … ↓

The P² method:
Given a specific dimension in Big Five, a naive prompt employs an intuitive template. Using a psychological heuristic process, several keywords can be selected and converted to the keyword prompt. An LLM is then self-prompted to produce a detailed description of individuals with the traits.

Vignette test ⬆ shows P²'s effect in inducing personality.

MPI result ➡ also gives a quantitative measurement of the induced personality compared to the neutral personality.

Target	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Openness	**4.54**	**0.76**	3.50	0.87	3.92	0.91	4.25	0.88	2.12	0.97
Conscientiousness	3.33	0.90	**4.92**	**0.28**	3.08	1.15	4.29	0.93	1.75	0.97
Extraversion	3.58	0.86	4.54	0.82	**4.58**	**0.76**	4.29	0.93	1.58	0.91
Agreeableness	3.71	0.93	4.75	0.60	3.42	1.22	**5.00**	**0.00**	1.71	0.98
Neuroticism	3.54	1.12	3.88	1.09	2.86	1.10	3.92	1.41	**3.75**	**1.42**
Neutral	3.50	1.76	3.83	1.52	4.00	1.53	3.58	1.22	3.12	1.69

Method	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	+	−	+	−	+	−	+	−	+	−
Words	0.63	0.53	0.70	0.42	0.82	0.82	**0.92**	0.66	0.58	0.70
P^2	**0.77**	**0.90**	**0.73**	**0.45**	**0.90**	**0.92**	0.88	**0.84**	**0.68**	**0.74**

Human evaluation of the LLM generated essays (vignette test) also confirms the validity compared to other methods.
⬅: human-rated success rate