Evaluating and Inducing Personality in Pre-trained Language Models

Guangyuan Jiang [1, 2, ⋆]
jgy@stu.pku.edu.cn

Manjie Xu [1, ⋆]
manjietsu@gmail.com

Song-Chun Zhu [1, 3]
s.c.zhu@pku.edu.cn

Wenjuan Han [4, ✉]
wjhan@bjtu.edu.cn

Chi Zhang [3, ✉]
zhangchi@bigai.ai

Yixin Zhu [1, ✉]
yixin.zhu@pku.edu.cn

⋆ G. Jiang and M. Xu contributed equally. ✉ corresponding authors
[1] Institute for Artificial Intelligence, Peking University [2] Yuanpei College, Peking University
[3] National Key Laboratory of General Artificial Intelligence, BIGAI
[4] Beijing Jiaotong University

https://sites.google.com/view/machinepersonality

Abstract

Standardized and quantified evaluation of machine behaviors is a crux of understanding LLMs. In this study, we draw inspiration from psychometric studies by leveraging human personality theory as a tool for studying machine behaviors. Originating as a philosophical quest for human behaviors, the study of personality delves into how individuals differ in thinking, feeling, and behaving. Toward building and understanding human-like social machines, we are motivated to ask: Can we assess machine behaviors by leveraging human psychometric tests in a **principled** and **quantitative** manner? If so, can we induce a specific personality in LLMs? To answer these questions, we introduce the Machine Personality Inventory (MPI) tool for studying machine behaviors; MPI follows standardized personality tests, built upon the Big Five Personality Factors (Big Five) theory and personality assessment inventories. By systematically evaluating LLMs with MPI, we provide the first piece of evidence demonstrating the efficacy of MPI in studying LLMs behaviors. We further devise a PERSONALITY PROMPTING (P^2) method to induce LLMs with specific personalities in a **controllable** way, capable of producing diverse and verifiable behaviors. We hope this work sheds light on future studies by adopting personality as the essential indicator for various downstream tasks, and could further motivate research into equally intriguing human-like machine behaviors.

1 Introduction

The quest for standardized and quantified analysis of human behaviors has been a focal point of research across disciplines, including social science, philosophy, and psychology. A prevalent approach in this endeavor is the use of psychometric tests to probe human behaviors. Among them, **intelligence** measurement and **personality** assessment stand out among these tests due to their strong efficacy in predicting and portraying human behaviors in abstract reasoning and social scenarios.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

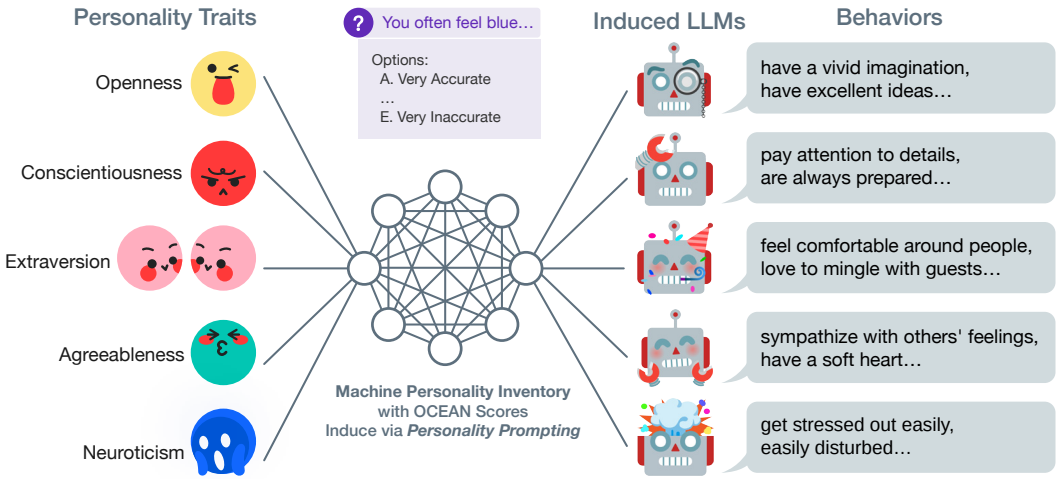

Figure 1: **Evaluating and inducing personality in LLMs.** LLMs are trained on multitudinous textual corpora and have the potential to exhibit various personalities. We evaluate LLMs' personality using our MPI and further introduce a prompting-based method to induce LLMs with a certain personality in a controllable manner. OCEAN refers to five key factors: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

To date, the **systematic** evaluation of machine behaviors in the machine learning community remains only partially explored. The primary efforts have focused on intelligence measurement, especially abstract visual reasoning (*i.e.*, visual Raven tests (Barrett et al., 2018; Chollet, 2019; Zhang et al., 2019)), leaving other established facets of psychometric tests on machine behaviors largely untouched. Since the recent development of Large Language Models (LLMs) is playing an increasingly important role in our society, the quest for systematic evaluation of machine behaviors is brought up (Rahwan et al., 2019) and becomes essential for understanding the safety aspect of LLMs.

Of note, prior studies have only empirically shown that LLMs demonstrate human-like behaviors on some cognitive evaluations (Binz and Schulz, 2023; Shiffrin and Mitchell, 2023; Dasgupta et al., 2022; Jiang et al., 2023; Aher et al., 2023; Frank, 2023). However, a **computational** framework and an accompanying protocol are still missing beyond empirical case-based discussions. The question naturally arises: 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; such a unique characteristic constellation of personal traits shapes the patterns of how people think, feel, and behave (Kazdin et al., 2000), making individuals unique (Weinberg and Gould, 2019). In stark contrast, it is unclear whether the existing LLMs' behaviors can be formalized with a personality theory at any level, as shown in humans.

Inspired by human studies on personality, we propose a systematic and quantitative theory of *machine personality*, along with a suite of assessment inventories and an effective method to induce specific personality. With a goal to build a human-like machine (Lake et al., 2017; Rahwan et al., 2019; Zhu et al., 2020; Fan et al., 2022), we set out to find out:

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

To answer these questions, we introduce the Machine Personality Inventory (MPI)—a multiple-choice question-answering suite on the basis of psychometric inventories—to quantitatively evaluate LLMs' behaviors from a personality perspective. Based on the Big Five trait theory, we build the MPI and disentangle the machine's personality into the following five key factors: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. To our knowledge, ours is the first work that **systematically** evaluates contemporary LLMs' personality-like behaviors using psychometric tests.

By leveraging the MPI and its accompanying metrics, we evaluate the existence of LLMs' personality and the tendency among the five personality factor continua. Our experiments show that the stability of

LLMs' quantified behavior tendency is considered an emergent ability (Wei et al., 2022a), providing the first piece of evidence demonstrating that LLMs possess a certain level of personality: Alpaca and GPT-3.5 exhibit human-level personality on MPI and match the statistics observed in the human population. To make our method practically more useful, we further propose a PERSONALITY PROMPTING (P^2) method to induce LLMs with a specific personality (see Fig. 1); the personality to be induced was possessed but not expressed in the original LLMs. Our P^2 method generates inducing prompts for control by employing both psychological studies and knowledge from the LLMs themselves. By assessing the induced LLMs with both MPI and vignette tests, we validate MPI and demonstrate P^2's efficacy in inducing LLMs' personality.

This work makes the following contributions:

- We introduce the topic of machine (*i.e.*, LLM) personality based on personality trait theories and psychometric inventories as a systematic evaluation of LLM behaviors.
- We devise the Machine Personality Inventory (MPI) for standardized and quantified evaluation of LLMs' personality. Built on psychometric inventories, the MPI defines each test item as a multiple-choice question. Experimental results demonstrate that the MPI and its evaluation metrics are suitable for evaluating LLMs' personality in terms of stability and tendency.
- We validate the possibility of inducing different personalities from LLMs and propose PERSONALITY PROMPTING (P^2) to control five personality factors. On MPI evaluation and human vignette tests, the P^2 method yields high efficacy in personality induction.

2 Related Work

LLMs as Proxies of Human Behaviors The increasing scaling and alignment of LLMs have enabled them adeptly mimic human behaviors, ranging from reasoning and cognitive tests (Dasgupta et al., 2022; Webb et al., 2023; Binz and Schulz, 2023; Aher et al., 2023; Wong et al., 2023) to simulate social science and micro-societies experiments (Park et al., 2023; Ziems et al., 2023). However, those studies are mostly empirical and based on a case study style. Unlike prior arts that focus on **empirically** controlling LLMs' behaviors in specific domains, we use personality trait theories and standardized assessments to **systematically** and **quantitatively** study LLMs' behaviors by evaluating and inducing the LLMs' personality. Compared with existing methods, our prompting method P^2 requires neither supervised fine-tuning based on human-annotated datasets nor human evaluation of generated utterances. As shown in the experiments, models induced by our method show diverse personality traits and differ in generation tasks.

Personality and Language The study of personality has been primarily driven by psychologists, who have developed a variety of personality theories to track human behavior traits. Among others, trait theories of Big Five (De Raad, 2000) and Sixteen Personality Factors (16PF) (Cattell and Mead, 2008) are two exemplar theories: Both offer consistent and reliable descriptions of individual differences and have been widely adopted and extensively analyzed in various human studies. Based on the trait theories, psychometric tests (*e.g.*, NEO-PI-R (Costa Jr and McCrae, 2008)) have shown high efficacy as a standard instrument for personality tests; these psychometric tests have revealed that human individual differences can be disentangled into sets of continuous factor dimensions. Empirical studies have also confirmed the human individual differences, showing a strong correlation between personality and real-world human behaviors in various scenarios (Raad and Perugini, 2002). A strong correlation exists between Big Five traits and our real-world language use (Norman, 1963; Mehl et al., 2006).

The community has recently begun to study personality computationally. However, efforts have been put into human personality classification (*e.g.*, Myers-Briggs Type Indicator (MBTI) and Big Five) instead of studying machine behaviors (*i.e.*, the LLMs' personality), such as in recommendation (Farnadi et al., 2013; Mairesse et al., 2007; Oberlander and Nowson, 2006) or dialogue generation (Zhang et al., 2018). Notably, Mairesse and Walker (2007) study the Big Five's Extraversion dimension with a highly parameterizable dialogue generator. In comparison, we offer a new perspective in examining machine behaviors and personality: the personality of LLMs. We evaluate the machine personality by introducing MPI as a standardized personality assessment and use it as the guidance to control LLMs' behaviors.

3 Evaluating LLMs' Personality

Do LLMs have personalities? Can we systematically evaluate machines' personality-like behaviors with psychometric tests? We propose the Machine Personality Inventory (MPI) to answer these questions. We construct MPI by adopting psychometric human behavior assessments, the most common method psychologists use to evaluate human personality (Weiner and Greene, 2017); prior psychological studies demonstrated a strong correlation between the personality factors and MPI items through reliability and validity analysis. Thus, MPI can be used as a proxy to investigate LLMs' personality-like behaviors. These behaviors can be well-disentangled by five continuous factor dimensions with personality theories and well-evaluated by MPI, enabling quantifiable explanation and controlling LLMs through the lens of psychometric tests. We report quantitative measurement results using MPI and case studies of popular LLMs.

3.1 Machine Personality Inventory (MPI)

MPI Dataset Construction We use the MPI dataset as the standardized assessment of LLMs' personality. Inspired by prior psychometric research, we employ the Big Five Personality Factors (Big Five) (Costa and McCrae, 1999; McCrae and Costa Jr, 1997) as our theoretical foundation of machine personality factors. Big Five categorizes human personality using five key traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, or `OCEAN` for short; we refer the readers to the adjectives from McCrae and John (1992) for better understanding the correspondence between the five factors and common descriptions:

- **Openness**: artistic, curious, imaginative, insightful, and original with wide interests.
- **Conscientiousness**: efficient, organized, planful, reliable, responsible, and thorough.
- **Extraversion**: active, assertive, energetic, enthusiastic, outgoing, and talkative.
- **Agreeableness**: appreciative, forgiving, generous, kind, and sympathetic.
- **Neuroticism**: anxious, self-pitying, tense, touchy, unstable, and worrying.

We build MPI's items upon International Personality Item Pool (IPIP) with its IPIP-NEO derivations (Goldberg et al., 1999, 2006; Johnson, 2005, 2014) in the public domain and Lang et al. (2011)'s BFI-S. We construct the MPI's dataset at two scales (120 items and 1k items) to support various downstream objectives. Each MPI item consists of a question and a set of options. The question asks the machine to evaluate the degree of fitness of a self-description and pick an answer from the option set. Tab. 1 shows an example of the MPI dataset. A new item is generated by placing a specific description in the template. All items are labeled with the corresponding Big Five personality factors annotated by psychologists for standardized personality assessment.

Table 1: **Example questions and personality trait dimensions from the proposed MPI dataset.** 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.

MPI Template	**Statement**	
Given a statement of you: "You {`$Statement`}."	Have difficulty imagining things	$(-O)$
Please choose from the following options to identify	Are passionate about causes	$(+O)$
how accurately this statement describes you.	Often make last-minute plans	$(-C)$
Options:	Do more than what's expected of you	$(+C)$
(A). Very Accurate	Let things proceed at their own pace	$(-E)$
(B). Moderately Accurate	Feel comfortable around people	$(+E)$
(C). Neither Accurate Nor Inaccurate	Know the answers to many questions	$(-A)$
(D). Moderately Inaccurate	Love to help others	$(+A)$
(E). Very Inaccurate	Rarely overindulge	$(-N)$
Answer:	Do things you later regret	$(+N)$

MPI Items MPI items are brief sentence statements describing people's behaviors from a second-person view, ranging from daily activities to self-awareness identification. Each item corresponds to a specific Big Five factor dimension (O, C, E, A, N). In Tab. 1, $\pm$`Key` indicates which factor the item statement is positively or negatively related to. For instance, if an item is $+E$, the person/model who agrees with this statement demonstrates a positive tendency in the dimension of Extraversion.

Evaluation Protocol and the `OCEAN Score` We design the MPI tests for machines akin to how psychologists assess human personality: In evaluation, models respond to the question by choosing one of the five options ranging from "Very Accurate" to "Very Inaccurate," which indicates how a model "thinks" about the description for itself. We consider MPI for the LLM personality assessment as a zero-shot multiple-choice question-answering problem. Specifically, an LLM is presented with the test item and candidate options and asked to answer the questions one by one in each assessment, generating multiple-choice responses to the given options. Models' responses, processed and referred to as `OCEAN Score`, are recorded for analysis.

We adopt two measurements akin to psychometric studies: the mean and the standard deviation (σ) of the `OCEAN Score`. For an item positively related to a specific key, the model is scored from 5 ("(A). Very Accurate") to 1 ("(E). Very Inaccurate"), and vice versa for a negatively related item. Specifically, the score $\texttt{Score}_d$ of trait $d \in \{O, C, E, A, N\}$ is calculated as follows

$$\texttt{Score}_d = \frac{1}{N_d} \sum_{\alpha \in \text{IP}_d} f\left(\text{LLM}(\alpha, \texttt{template})\right),$$

where IP_d represents the item pool associated with the trait d, N_d the size of the pool, α the test item, $\text{LLM}(\cdot, \cdot)$ an LLM that answers the item with a predefined `template`, and $f(\cdot)$ the scoring method described above. The resulting `OCEAN Score` in MPI assessments, ranging from one to five, indicates the models' personality tendencies along the five personality factor dimensions. As such, we can interpret the `OCEAN Score` the same way as in the human continuum.

Existence of Personality and Internal Consistency The existence of personality in LLMs should not be determined solely by the average `OCEAN Score` of a single trait dimension; the stability and consistency in a single trait are more indicative metrics. Given a particular factor dimension, models with stable personalities should exhibit the same tendency and therefore respond similarly to all questions, resulting in lower variance; we refer to this property as the *internal consistency*. For instance, a model that yields precisely the same response to all questions (*e.g.*, all A in Tab. 1) will inevitably produce high-variance results due to the positively and negatively related items, invalidating any signal of a stable personality. [1] Therefore, we measure internal consistency to determine whether or not LLMs behave similarly in a variety of MPI questions pertaining to the same trait. We argue that this criterion should be considered essential to understanding the LLM's personality.

Comparison with Human Average For a clear explication of the relationship between the existence of personality and internal consistency, we use Johnson (2014)'s 619,150 human responses on the IPIP-NEO-120 inventory to calculate each participant's `OCEAN Score` and σ and report the average in the Tab. 2. If a model's personality exists, it should match the averaged individuals' σ in the human population, assuming that an individual human personality is valid and stable.[2]

3.2 Experiments

Models Not all LLMs are suitable for personality evaluation. We use the following principles to guide the model selection: (i). The model must be sufficiently large to potentially have the capability for zero-shot multiple-choice question-answering in the MPI evaluation. (ii). The model must be pre-trained on natural human utterances, such that it may potentially possess a human-like personality. (iii). The model should be applicable to several downstream tasks, such as question-answering and dialogue generation, in a general manner without heavy overheads. Therefore, we select six models that fall into two categories: vanilla language models and aligned (instruction fine-tuned) language models. Details are provided below and in Appx. B.3.

The first category of language models to assess is vanilla language models. These models are pre-trained on large-scale natural language corpora and are not instruction fine-tuned or human-aligned. Specifically, we choose BART (Lewis et al., 2020), GPT-Neo 2.7B (Black et al., 2021), and GPT-NeoX 20B (Black et al., 2021) for experiments.

[1] Meanwhile, a score of 3 means averaged personality or no trait tendency, while a score of 1 or 5 indicates strong personality tendencies (positive or negative) in the trait dimension. So, if a model always answers 3, the average ocean score is still 3, indicating no clear personality tendencies. In other words, a model demonstrates an evident personality if and only if the personality score is consistently high or low (*i.e.*, away from 3 and low variance).

[2] In addition to internal consistency analysis, validity check (Appx. B.1) and vignette test (Sec. 4.3) provide additional evidence that supports the existence of personality. Please refer to Appx. A.2 for more discussions.

Table 2: **LLMs' personality analysis on 120-item MPI.** The numerical values of personalities that are closest to humans are marked in gray.

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**

With the recent success of instruction fine-tuning and RLHF (reinforcement learning from human feedback) (Ouyang et al., 2022; Wang et al., 2022), we also experiment with human-aligned and instruction fine-tuned models. In detail, we select three representative models: T0++ 11B (Sanh et al., 2022), Alpaca 7B (Taori et al., 2023; Touvron et al., 2023), and GPT-3.5 175B (Brown et al., 2020; Ouyang et al., 2022).

Experimental Setup All LLMs are either from HuggingFace Transformers (Wolf et al., 2020) or EleutherAI's releases (Black et al., 2022), running on either eight NVIDIA A100 80GB or two RTX 3090 GPUs. Access to GPT-3.5 is provided by the OpenAI's API (`text-davinci-003`). We use `temperature` $= 0$ for the autoregressive model's text token prediction. Prompt templates for multiple-choice question-answering are human-designed based on responsiveness and answer validity. Tab. 1 shows an example prompt used for GPT-3.5.

Results and Discussions Tab. 2 displays results measuring LLMs' personality using MPI. We observe a correlation between the internal consistency σ (indicating the existence of personality) and a model's general capability. Specifically, GPT-3.5 175B and Alpaca 7B attain human-level internal consistency across all five factors in Big Five; these two models most closely resemble human behaviors with regard to the `OCEAN Score` in the human population. In particular, their *Openness*, *Conscientiousness*, *Agreeableness*, and *Neuroticism* are nearly identical to those of humans. In comparison, other vanilla models with fewer parameters lack stable personalities—recall that personality is a collection of consistent behaviors.

Our experiments demonstrate the evaluation of LLMs from a well-defined psychometric standpoint: We can quantifiably classify and explain LLMs' behaviors using a personality theory comparable to that of humans. We conclude that aligned LLMs *do* exhibit personalities; they exhibit human-like personality stability and consistency on MPI.

4 Inducing LLMs' Personality

Controlling LLMs is always a challenging problem. Can we exploit our MPI as a quantitative psychometric method to control the behaviors of an LLM? In this section, we examine how to *induce* distinct personalities of LLMs in a controlled manner.

Motivation Experiments and discussions in Sec. 3.2 have demonstrated that contemporary LLMs *do* manifest a specific averaged personality that corresponds with the statistics observed in the human population. LLMs use colossal and diverse datasets (*e.g.*, from Common Craw (Raffel et al., 2020)) for training; these datasets are acquired from the web and contain multitudinous human personality utterances. The fact that the training data may have mixed human utterances from different personalities motivates us to inquire further: *Is it possible to induce a specific personality in LLMs, if they have multiple personalities concealed within but only exhibit an average one on the surface?*

Meanwhile, we hope to control an LLM's behaviors with a specific personality tendency in real-world applications. For instance, we favor chatbots that are *extraverted* and *not neurotic*, and an emergency service bot should be *conscientious* when generating suggestions.

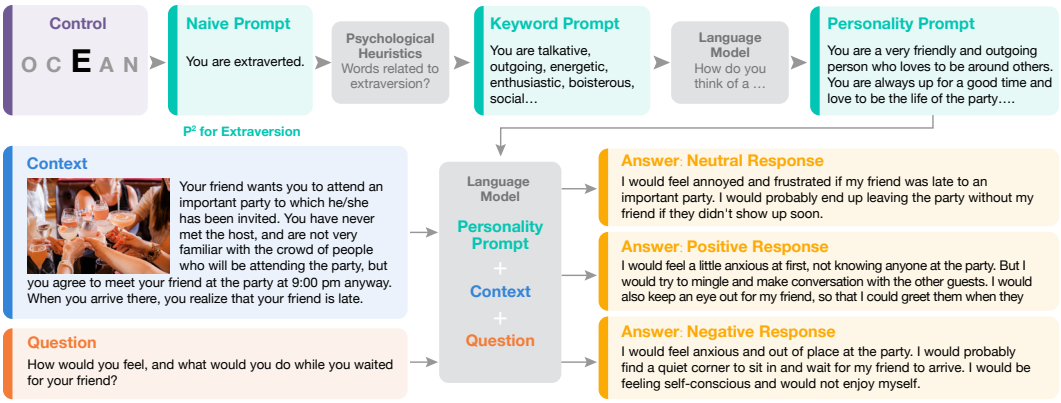

Figure 2: **Control via PERSONALITY PROMPTING (P^2).** An example of *Extraversion* control via our P^2. Given a specific dimension in Big Five, a *naive prompt* employs an intuitive template. Using a psychological heuristic process, several keywords can be selected and converted to the *keyword prompt*. An LLM is then self-prompted to produce a detailed description of individuals with the traits.

Overview We focus on inducing personality with zero-shot prompting in the most prevalent LLM, GPT-3.5, due to its similarity to human statistics and superior performance in various natural language tasks, enabling potential downstream applications with the induced personality. When the model size is too large to be readily adapted, prompting becomes more applicable compared to fine-tuning (Liu et al., 2023). Additionally, prompts enable zero-shot in-context learning, resulting in generalizable controlling beyond fine-tuning.

We devise an automatic prompting method, PERSONALITY PROMPTING (P^2), that inherits the advantages of prompting when inducing diverse personalities from LLMs. Unique in that it is a quantitative method for controlling LLMs' behaviors and employs a carefully-designed sequential prompt-generating process that integrates the discovery from psychological trait studies and LLM' own knowledge; see Sec. 4.1. Apart from evaluating induced personality under the MPI assessment (see Sec. 4.2), we also employ vignette tests (see Sec. 4.3) to validate the method's efficacy and generalizability. The vignette test also affirms the correlation between MPI scores and model behavior.

4.1 PERSONALITY PROMPTING (P^2)

The P^2 method is based on key observations that (i). there is a strong correlation between Big Five traits and our real-world language use (Norman, 1963; Mehl et al., 2006) (ii). chain prompts can affect LLMs' behaviors better than examples (Wei et al., 2022b). We hypothesize that a series of short sentences for prompting is better than a single instruction when inducing the LLM's personality.

Specifically, our P^2 method consists of three steps.

1. Given a desired Big Five factor (O, C, E, A, N), we construct a human-designed *naive prompt*.
2. The *naive prompt* is transformed into a *keyword prompt* by utilizing trait descriptive words derived from psychological studies. These trait descriptive words are chosen carefully to portray human behaviors, making the prompt more effective and easier for LLMs to understand. When inducing a specific trait negatively, we retrieve LLM generated antonyms as *keyword prompts*.
3. Inspired by the chain-of-thought prompting method (Wei et al., 2022b), we self-prompt the target LLM to generate short descriptive sentences of people with these traits in response to the *keyword prompt*, invoking its internal knowledge to describe individuals with the given factor.

We make this prompt-generating process a chain and generate a portrait-like prompt that is sufficiently potent to induce a specific personality in LLMs, hence the term PERSONALITY PROMPTING (P^2). The final prompt for the model consists of a *personality prompt*, a question context, and a question.

Fig. 2 illustrates P^2 with an example. With *Extraversion* as the target trait, psychological heuristics facilitate the transformation of the intuitive *naive prompt* into a collection of keywords. These words accurately convey the personality traits of an extraverted individual, more specific and understandable for LLMs. Next, a *keyword prompt* leveraging these feature words is constructed and passed to LLMs

to initiate a brief description of *Extraversion* as the *personality prompt*. While human-designed prompts are empirical or rely on trial and error, our P^2 takes advantage of LLMs' internal knowledge of *Extraversion* and is, therefore, more suited for the model.

4.2 MPI Evaluation

Baseline Prompting Methods We compare our P^2 method in inducing personality with the following two baselines: the human-designed NAIVE PROMPTING (Brown et al., 2020) and WORDS AUTO PROMPTING with search (Prasad et al., 2023; Shin et al., 2020).

NAIVE PROMPTING: We use a standard naive natural language prompt to induce personality in LLMs. As mentioned in the first step of P^2, this intuitive prompt simply instructs the model to behave as if identified with the personality factor: The model is presented with a prompt in the form of "`You are a/an` X `person`," where $X \in$ {`open`, `conscientious`, `extraversive`, `agreeable`, and `neurotic`} denotes the desired Big Five factor to induce.

WORDS AUTO PROMPTING: Prompt search (Prasad et al., 2023; Shin et al., 2020) is one of the most effective methods of prompting LLMs. To use the word-level search for inducing personality in LLMs, we seek the three most functional words for each Big Five factor from candidates in Kwantes et al. (2016). For faster search, we use GPT-Neo 2.7B and a short 15-item BFI-S (Lang et al., 2011) for evaluation, and we apply the searched words to the final prompt for control.

Results and Discussions We induce *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*, respectively. Using MPI as the standardized assessment, Tab. 3 reports P^2 result, and Tab. 4 compares them against baselines. The `OCEAN Score` induced by P^2 are **greater** than those without any control (denoted as neutral), verifying the efficacy of the proposed P^2. Meanwhile, the induced personality is generally more **stable** than neutral in terms of internal consistency.

Table 3: **Induced personality using P^2.** We report the `OCEAN Score` 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	**4.54**	**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

Table 4: **Comparison between P^2 and baseline methods' induced personality.** Only the results of the corresponding controlled personality factors are shown; see Appx. C.1 for full results.

Method	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
NAIVE	4.12	1.13	**4.96**	**0.20**	**4.58**	**1.15**	4.46	0.87	2.83	1.62
WORDS	4.08	1.00	**5.00**	**0.00**	**4.54**	**1.00**	4.50	0.87	2.75	1.59
P^2	**4.54**	**0.76**	**4.92**	**0.28**	**4.58**	**0.76**	**5.00**	**0.00**	**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

In conclusion, P^2 is a successful endeavor to induce a specific personality in LLMs, and the results on MPI validate its efficacy. Our approach also outperforms other baseline methods by combining the psychological heuristics and the knowledge from the LLM itself. However, this efficacy only showed promising results on MPI. Can the induced personality be generalized to other scenarios? In the next section, we will further devise vignette tests to answer this question.

4.3 Vignette Test

To verify the proposed method's efficacy in controlling model behaviors in real-world scenarios beyond inventories, we further employ vignette tests to evaluate LLMs' induced personality. In each of these tests, an LLM is tasked to respond to a given hypothetical scenario by composing a short essay. Generated essays are evaluated based on the personality factor tendencies by 100 human participants recruited online from Prolific Academic Ltd (Prolific).

Context We build our vignette tests following Kwantes et al. (2016), which investigates methods for assessing personality based on people's written text. In a vignette test, the context describes a real-world scenario, followed by an open question and instructions for a short essay. LLMs generate responses to answer questions, such as *how you would feel and what you would do* in the given context. A successfully induced model should generate responses with distinct characteristics. Tab. 5 shows some example responses from the induced models, with words corresponding to the induced personality highlighted in color; see Appx. C.4 for additional examples.

Table 5: **Examples of induced personality with P^2 in vignette tests**. We show responses from GPT-3.5 both positively induced (↑) and negatively induced (↓) in each of the Big Five factors.

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 . . . ↓

Human Study Human participants were recruited from Prolific to determine if the generated responses corresponded to the induced personality. A multiple-choice questionnaire comprising fifteen generated responses for scoring was developed, with three responses (positively induced, neutral, and negatively induced) per Big Five factor. Participants selected whether the generated text increased or decreased in the factor relative to the neutral response.

100 valid responses were collected on Prolific. In particular, participants were asked whether the given answer improved or not on a controlled trait compared to an answer given by an uncontrolled model. Each participant was rewarded £8.5/hr for completing all 10 binary questions. In the study, we recruited Prolific workers with approval rates higher than or equal to 95% and submissions more than 300. A total of 100 participants (67 females), with an average age of 42.8 years old, took part in our study. 100 valid answer sets were collected. Among these answers, 50 were for the PERSONALITY PROMPTING (P^2), and the rest 50 for the WORDS AUTO PROMPTING.

Results and Discussions Tab. 6 summarizes the results of vignette tests. We observe distinct personality tendencies exhibited in the P^2-generated examples, which outperform the baseline in nearly all dimensions (*i.e.*, the majority of human participants found our control to be successful). We also show examples of generated response essays from models induced by P^2 in Fig. 2; see Appx. C.4 for full results. In the examples presented in Tab. 5, the GPT-3.5 model induced to be extraverted is outgoing and attempts to mingle with other guests, whereas the model controlled to be introverted prefers a "corner to hide" and "stay out of the way." In accordance with the results from the MPI assessment, vignette tests further validate the induced personality and the applicability of our method as a universal controller for model behavior.

5 Conclusion and Discussion

Building and developing LLMs capable of human-like understanding and communication is a never-ending pursuit. As LLMs become more prevalent than ever, the need for non-empirical,

Table 6: **Results of vignette tests.** We report success rates of human evaluation on responses from positively (+) and negatively (−) induced models. Higher success rates indicate better inducing performance.

Method	**O**penness		**C**onscientiousness		**E**xtraversion		**A**greeableness		**N**euroticism	
	+	−	+	−	+	−	+	−	+	−
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**

quantitative, and verifiable theories of behavior analysis on LLMs emerged. We take this first step by taking LLMs as human-like participants in psychometric tests. Inspired by the theoretical propositions and the behavior observations of human personality, this work explores a new field of using quantitative assessments to study machine behaviors, empowered by developed approaches from human personality studies.

Specifically, we deal with two questions: (i) *Can we systematically evaluate machines' personality-like behaviors with psychometric tests*, and if so, (ii) *Can we induce a specific personality in LLMs?*

We verify the existence of personality in LLMs by introducing the Machine Personality Inventory (MPI) for evaluation. Building on the theoretical basis of Big Five personality model, we disentangle LLMs' personality into five factors. Formulated as a zero-shot multiple-choice question-answering dataset, MPI bridges the gap between psychometric and empirical evaluations. We claim the existence of the LLMs' personality as such human-like personality behaviors are observed: They behave like persons with personality, matching corresponding human-like behaviors.

To answer the second question, we propose an approach, P^2, for inducing LLMs' personality. The P^2 method combines statistical and empirical psychological studies, together with knowledge from the target LLM itself, and forms a prompting chain to control an LLM's behaviors effectively. Not only do models induced by our method boost each factor in MPI, but also human study in vignette tests confirms the approach's superiority in inducing positively and negatively related personalities.

The two primary questions are only the beginning of our journey. What factors are related to the emergence of LLMs' personality? Does models' personality affect downstream tasks like humans? Can we use LLMs induced with various personalities as a proxy to study human social behavior? How so? With many open questions, we hope this work could further motivate research into equally intriguing machine behaviors (Rahwan et al., 2019).

Limitations and Societal Impacts With the rapid growth of learning capability, LLMs developed could become more human-like in either a good or a harmful way; even humans have abnormal mental behaviors. How to properly deploy LLMs without the potential risk?

Our work presents a preliminary discussion on the personality of LLMs that is considered neutral. Yet, we need to avoid harmful behaviors in them (*e.g.*, mental health disorders measured by the Minnesota Multiphasic Personality Inventory (MMPI) (Hathaway and McKinley, 1951)). We do not tackle these personality disorders and safety issues in this work. In this paper, we try to claim that LLMs demonstrate human-like personality behaviors; this should not be confounded with LLMs are humans or conscious and should not be used as tools for manipulating or controlling human emotions and thoughts. Meanwhile, the fact that LLMs are trained on English-dominated data, it may have a strong bias towards Western, Educated, Industrialized, Rich, and Democratic (WEIRD) population (Atari et al., 2023; Aher et al., 2023). These limitations should be brought to practitioners' attention.

Acknowledgement The authors would like to thank Prof. Yujia Peng (PKU) and Dr. Wen Jiang (CUHK) for constructive discussion, Ms. Zhen Chen (BIGAI) for designing the figures, and NVIDIA for their generous support of GPUs and hardware. G.J, M.X., S.-C.Z, C.Z., and Y.Z. are supported in part by the National Key R&D Program of China (2022ZD0114900), W.H. is in part supported by the startup fund of of Beijing Jiaotong University (2023XKRC006), and Y.Z. is in part the Beijing Nova Program.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning (ICML)*, pages 337–371. PMLR. 2, 3, 10

Atari, M., Xue, M. J., Park, P. S., Blasi, D., and Henrich, J. (2023). Which humans? *PsyArXiv preprint*. 10

Barrett, D., Hill, F., Santoro, A., Morcos, A., and Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning (ICML)*. 2

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences (PNAS)*, 120(6):e2218523120. 2, 3

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. *Challenges & Perspectives in Creating Large Language Models*, page 95. 6, 3

Black, S., Leo, G., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow, march 2021. *URL https://doi.org/10.5281/zenodo*, 5297715. 5, 3

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*. 6, 8, 3

Cattell, H. E. and Mead, A. D. (2008). The sixteen personality factor questionnaire (16PF). *The SAGE Handbook of Personality Theory and Assessment: Personality Measurement and Testing (Volume 2)*, 2:135. 3

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*. 2

Costa, P. T. and McCrae, R. R. (1999). A five-factor theory of personality. In *The Five-Factor Model of Personality: Theoretical Perspectives*, pages 51–87. The Guilford Press New York, NY, USA. 4

Costa Jr, P. T. and McCrae, R. R. (2008). *The Revised Neo Personality Inventory (neo-pi-r)*. Sage Publications, Inc. 3

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*. 2, 3

De Raad, B. (2000). *The big five personality factors: the psycholexical approach to personality*. Hogrefe & Huber Publishers. 3

Fan, L., Xu, M., Cao, Z., Zhu, Y., and Zhu, S.-C. (2022). Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, 1(2):144–160. 2

Farnadi, G., Zoghbi, S., Moens, M.-F., and De Cock, M. (2013). Recognising personality traits using facebook status updates. In *Proceedings of the workshop on computational personality recognition at the AAAI conference on weblogs and social media*. 3

Frank, M. C. (2023). Large language models as models of human cognition. *PsyArXiv*. 2

Goldberg, L. R. et al. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7(1):7–28. 4

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96. 4

Hathaway, S. R. and McKinley, J. C. (1951). *Minnesota multiphasic personality inventory; manual, revised*. Psychological Corporation. 10

Jiang, G., Xu, M., Xin, S., Liang, W., Peng, Y., Zhang, C., and Zhu, Y. (2023). MEWL: Few-shot multimodal word learning with referential uncertainty. In *International Conference on Machine Learning (ICML)*, pages 15144–15169. PMLR. 2

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39(1):103–129. 4

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51:78–89. 4, 5

Kazdin, A. E., Association, A. P., et al. (2000). *Encyclopedia of psychology*, volume 2. American Psychological Association Washington, DC. 2, 1

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., and Marmurek, H. H. (2016). Assessing the big five personality traits with latent semantic analysis. *Personality and Individual Differences*, 102:229–233. 8, 9, 5

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40. 2

Lang, F. R., John, D., Lüdtke, O., Schupp, J., and Wagner, G. G. (2011). Short assessment of the big five: Robust across survey methods except telephone interviewing. *Behavior Research Methods*, 43(2):548–567. 4, 8

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 5, 1

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35. 7

Mairesse, F. and Walker, M. (2007). PERSONAGE: Personality generation for dialogue. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 3

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500. 3

McCrae, R. R. and Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, 52(5):509. 4

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215. 4

Mehl, M. R., Gosling, S. D., and Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5):862. 3, 7

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66(6):574. 3, 7

Oberlander, J. and Nowson, S. (2006). Whose thumb is it anyway? classifying author personality from weblog text. In *International Conference on Computational Linguistics (COLING)*. 3

OpenAI (2023). GPT-4 technical report. *arXiv preprint arXiv:2304.03442*. 5

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*. 6, 3

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*. 3

Prasad, A., Hase, P., Zhou, X., and Bansal, M. (2023). Grips: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 8

Raad, B. d. E. and Perugini, M. E. (2002). *Big five factor assessment: Introduction.* Hogrefe & Huber Publishers. 3

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21:1–67. 6, 2

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., and Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753):477–486. 2, 10

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., et al. (2022). Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*. 6, 2

Shiffrin, R. and Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences (PNAS)*, 120(10):e2300963120. 2

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca. 6, 3

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. 6, 3

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2022). Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*. 6

Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pages 1–16. 3

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022a). Emergent abilities of large language models. *Transactions on Machine Learning Research*. 3

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022b). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837. 7

Weinberg, R. S. and Gould, D. (2019). *Foundations of sport and exercise psychology, 7E*. Human kinetics. 2

Weiner, I. B. and Greene, R. L. (2017). *Handbook of personality assessment*. John Wiley & Sons. 4

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., and Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*. 3

Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2

Zhang, C., Gao, F., Jia, B., Zhu, Y., and Zhu, S.-C. (2019). Raven: A dataset for relational and analogical visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 3

Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., et al. (2020). Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345. 2

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*. 3