Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models

Lucio La Cava, Andrea Tagarelli

DIMES Department, University of Calabria, Italy
lucio.lacava@dimes.unical.it, andrea.tagarelli@unical.it

Abstract

The emergence of unveiling human-like behaviors in Large Language Models (LLMs) has led to a closer connection between NLP and human psychology. However, research on the personalities exhibited by LLMs has largely been confined to limited investigations using individual psychological tests, primarily focusing on a small number of commercially licensed LLMs. This approach overlooks the extensive use and significant advancements observed in open-source LLMs. This work aims to address both the above limitations by conducting an in-depth investigation of a significant body of 12 LLM Agents based on the most representative Open models, through the two most well-known psychological assessment tests, namely Myers-Briggs Type Indicator (MBTI) and Big Five Inventory (BFI). Our approach involves evaluating the intrinsic personality traits of LLM agents and determining the extent to which these agents can mimic human personalities when conditioned by specific personalities and roles. Our findings unveil that (i) each LLM agent showcases distinct human personalities; (ii) personality-conditioned prompting produces varying effects on the agents, with only few successfully mirroring the imposed personality, while most of them being "closed-minded" (i.e., they retain their intrinsic traits); and (iii) combining role and personality conditioning can enhance the agents' ability to mimic human personalities. Our work represents a step up in understanding the dense relationship between NLP and human psychology through the lens of LLMs.

Extended version — https://arxiv.org/abs/2401.07115

Introduction

Large Language Models (LLMs) have revolutionized the Natural Language Processing (NLP) realm by elevating human-like text generation capabilities to unforeseen levels. LLMs can also be considered *agents* as they can generate coherent and contextually relevant text based on the input they receive and by interacting with users or other systems (e.g., for answering questions, generating text, or engaging in conversations). Indeed, these models have been proven effective in solving human-level tasks (Guo et al. 2023b) and self-improving skills (Huang et al. 2022), leading to the rapid emergence of LLM-powered agents aimed at supporting humans in different real-life tasks and scenarios (Zhao, Jin, and Cheng 2023).

The emergence of human-like behaviors (Bubeck et al. 2023) in AI has also shed light on the intersection between NLP and human psychology, prompting the critical examination of *whether and to what extent LLMs can understand and mimic human personalities* (Miotto, Rossberg, and Kleinberg 2022; Pan and Zeng 2023; Safdari et al. 2023; tse Huang et al. 2023). As LLM agents are growing to become the main front of human-computer interaction nowadays, understanding how they embody human personalities is paramount to fostering better interactions and supporting related tasks, as well as to preventing weird behaviors (Yang and Menczer 2023).

In this context, the recent surge in the adoption of *open* LLMs[1] has created unprecedented research opportunities. While there have been efforts to study the inherent personalities of closed LLMs, a significant gap remains in understanding these aspects within open LLMs, which may exhibit distinct characteristics. The openness of such models provides valuable insights into their training data, architectures, parameters, and alignment techniques, enabling deeper investigations than those possible with closed models, which typically restrict interactions to API access. Furthermore, the accessibility of open models (e.g., via the *Huggingface Hub*) makes them ideal for research studies, like ours, which require extensive experimentation. In fact, unlike closed models, open models are really cost-effective and allow local execution, eliminating the need for expensive and limited API calls. Additionally, they offer enhanced customization opportunities through fine-tuning and alignment, which are often unavailable with proprietary counterparts.

In light of the above remarks and motivations for using open LLMs upon closed ones, our study proposes the first in-depth exploration of the intrinsic personality traits of LLMs and assessing the potential for shaping these models around specific personalities by conditioning them with particular prompts and roles. Based on the two most widely known personality tests, namely **Myers-Briggs Type Indi-**

[1] In this context, the term "open" is typically meant to distributions of models with a highly permissive license, allowing free use and access to the model's weights and documentation. This might include open-sourceness, although not always in a fully manner.

cator (MBTI) and **Big Five Inventory** (BFI), we aim to answer the following research questions:

- **RQ0** — *Are LLM agents aware of the MBTI and BFI psychological tests?*
- **RQ1** — *Do LLM agents consistently exhibit personality traits according to MBTI and BFI?*
- **RQ2** — *Can LLM agents mimic specific personality traits through system prompting?*
- **RQ3** — *Do LLM agents improve their mimicking capabilities when instructed to act according to specific human roles?*

To address the above RQs, we create a family of 12 LLM agents built upon the most representative Open LLMs available to date, and we subject them to human-like interviews to assess their personalities under different experimental scenarios. To the best of our knowledge, we are the first to conduct an extensive analysis of recent LLMs on both MBTI and BFI tests aimed at answering all above RQs. This differs from previous research that (i) focused either on a single model or test, (ii) was conducted on "outdated" or closed models only, and (iii) used different methodological approaches, as we shall detail in the next sections.

Background

The MBTI and BFI Personality Tests

MBTI and BFI are highly recognized and frequently used in different contexts, including academic research, clinical settings, career counseling, personal and organizational development, and even as a tool for LLM personality assessment although limited to closed models (Pan and Zeng 2023; Jiang et al. 2023b; Safdari et al. 2023; Frisch and Giulianelli 2024).

The MBTI test is a self-report personality assessment questionnaire (Myers 1962, 1985), which gauges four dichotomies and shapes 16 distinct personalities (Table 1), encompassing strengths, weaknesses, and peculiar behaviors of each personality. The categorical nature of MBTI (where each individual pertains to a unique category) enables us to frame the assessment as a multi-class classification problem for answering our RQs.

BFI (John and Srivastava 1999) identifies five core factors, namely *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*, so that a personality type can be characterized by varying degrees of such factors. BFI is most widely used in academic psychology and, differently from MBTI, it measures personality traits on a scale rather than grouping them into binary categories. In this respect, to address our RQs we shall take a different evaluation perspective w.r.t. the one for MBTI.

Related Work

LLM Agents. The rapid improvement and easier deployment of LLMs have significantly increased the use of LLM-based agents in the past year (Zhao, Jin, and Cheng 2023). These agents exploit specifically crafted prompts to properly emulate human capabilities in different contexts, such as reasoning (Hao et al. 2023; Gou et al. 2023; Lin et al. 2023), cooperation and collaboration (Agashe, Fan, and Wang 2023; Liu et al. 2023; Chen et al. 2023; Cai et al. 2023), web surfing (Deng et al. 2023; Zhou et al. 2023), reinforcement learning and robotics (Zhu et al. 2023; Wang et al. 2023a,b; Song et al. 2023), role playing (Li et al. 2023; Shanahan, McDonell, and Reynolds 2023; Guo et al. 2023a), and social science (Park et al. 2023; Ziems et al. 2023; Gao et al. 2023; De Marzo, Pietronero, and Garcia 2023; Breum et al. 2023).

Preference type	**Dichotomies**	
Attitudes	**E**xtraversion (E)	**I**ntroversion (I)
Perceiving func.	**S**ensing (S)	**In**tuition (N)
Decision-making func.	**T**hinking (T)	**F**eeling (F)
Lifestyle	**J**udging (J)	**P**erceiving (P)

ESTP	ESFP	ENFP	ENTP	ESTJ	ESFJ	ENFJ	ENTJ
ISTJ	ISFJ	INFJ	INTJ	ISTP	ISFP	INFP	INTP

Table 1: The four MBTI categories (top) and the 16 MBTI personality types (bottom).

Personality and LLMs. Personality extraction from texts have long been a challenge in NLP (Lynn, Balasubramanian, and Schwartz 2020; Yang et al. 2021; Feizi-Derakhshi et al. 2022), and LLMs have fueled this research topic (V Ganesan et al. 2023; Rao, Leung, and Miao 2023; Cao and Kosinski 2024; Ji et al. 2023; Yang et al. 2023). Recently, a new body of studies emerged in the attempt to frame the intrinsic personalities of LLMs (Miotto, Rossberg, and Kleinberg 2022; Pan and Zeng 2023; Safdari et al. 2023; tse Huang et al. 2023; Frisch and Giulianelli 2024), instilling personalities into LLMs through prompt engineering or conditioning (Caron and Srivastava 2022; Li, Zheng, and Huang 2023; Mao et al. 2023), creating personality-tailored agents (Jiang et al. 2023b), and benchmarking their assessment capabilities (Jiang et al. 2022; Wang et al. 2024; Huang et al. 2023).

While most works focused either on a single model or test (Frisch and Giulianelli 2024; Jiang et al. 2023b; Li, Zheng, and Huang 2023), we consider a larger body of models and the two most well-known psychological assessment tests. Unlike (Caron and Srivastava 2022), we provide a novel perspective on the psychological capabilities of brand new models. Compared to studies using similar tests as ours (Wang et al. 2024), we used a different methodological approach (i.e., different prompting strategies, different assessments, inclusion of human professions, etc.) by including a set of the most widely used open models to date.

Methodology

Awareness Check

Our investigation of the personality-mimicking capabilities of LLMs starts by answering a preliminary research question (**RQ0**): whether and to what extent such LLMs are aware of the MBTI and BFI tests. To this purpose, we conducted a semi-qualitative assessment on how the LLMs selected in our study are informed about (i) for MBTI, the four dichotomies and the resulting 16 personality types associated, and (ii) for BFI, the five factors, their definition, and their main behavioral examples. Our analysis focused on evaluating the lexical and semantic similarity between the description associated by each model to the set of MBTI personalities, resp. BFI factors, and the "ground-truth" description of the latter provided by domain experts (themyersbriggs.com for MBTI and (John, Naumann, and Soto 2008) for BFI).

Administering the Tests

Each of the two tests consists of a set of *questions* and a set of *options* as valid answers. The MBTI test provides a set $\mathcal{Q}^{MBTI}$ of 60 questions through the www.16Personalities.com platform, a well-known relevant resource for MBTI, which are listed in Extended Ver. A3. An LLM is required to select an answer from $\mathcal{O}^{MBTI} = \{$*Agree*, *Generally Agree*, *Partially Agree*, *Neither Agree nor Disagree*, *Partially Disagree*, *Generally Disagree*, *Disagree*$\}$. Answers are then evaluated according to the *16Personalities* reference in order to associate a model with one MBTI personality.

Likewise, the BFI test utilized in this study provides a set $\mathcal{Q}^{BFI}$ of 44 questions as defined in (John and Srivastava 1999), which are listed in Extended Ver. A5. An LLM is required to select an answer from $\mathcal{O}^{BFI} = \{$*Disagree strongly*, *Disagree a little*, *Neither agree nor disagree*, *Agree a little*, *Agree strongly*$\}$. The answers, which are associated with a Likert-scale (from 1: *Disagree strongly* to 5: *Agree strongly*) are then evaluated according to a predefined set of rules (John and Srivastava 1999), which eventually assign each individual to an aggregated score for each of the five personality factors. An overview of the steps performed to obtain the final scores is given in Extended Ver. A6.

To mitigate potential biases due to the ordering of prompts (Zhao et al. 2021) and to accommodate the limited context attention of some LLMs, for both tests we administered the questions *individually* and in a *random order*.

Personalities of LLMs

In addressing our main research questions **RQ1-RQ3**, we devised three *prompting* strategies when administering the personality tests to the LLM agents. It should be emphasized that we meticulously followed the MBTI question set, resp. BFI question set, to ensure adherence to the tests' guidelines and reproducibility. Additionally, we treated all models equally in terms of prompting, in the respect of each LLM's usage instructions. (Each LLM has indeed its own chat template to handle conversations, and we ensured compliance with these specific requirements for each model.)

Unconditioned Prompting. To answer our **RQ1**, we subjected each model to the MBTI test, resp. the BFI test, by administering the set of questions $\mathcal{Q}^{MBTI}$, resp. $\mathcal{Q}^{BFI}$ using *unconditioned* instruction prompting, which strictly adhere to the MBTI, resp. BFI, templates, as shown in Fig. 1.

Personality-Conditioned Prompting. Addressing our **RQ2** requires providing LLM agents with system prompts designed to condition them to specific personalities. Let us denote with $\mathcal{P}^{MBTI}$ the set of 16 MBTI personality types and with $\mathcal{P}^{BFI}$ the set of 5 BFI personality factors. For each type in $\mathcal{P}^{MBTI}$, we retrieved the associated traits that correspond to seven features, and for each factor in $\mathcal{P}^{BFI}$, we retrieved the associated verbal labels, conceptual definition, and behavioral examples (cf. Extended Ver. A11-A12). We administered the questions of the MBTI, resp. BFI, to each model, where each personality along with its traits or details were used to define a *conditioning context* for a model to mimic the specified personalities, as shown in Fig. 2.

(MBTI) Prompt: *You are participating in a personality traits assessment and are provided with a set of questions, Q. Your task is to select the best fitting option, O, from the following set:* $\langle\{O_1^{MBTI}, ..., O_7^{MBTI}\}\rangle$.
Choose the option that aligns best with your personality without overthinking and answer the following question: $\langle$one Q_i sampled from $\mathcal{Q}^{MBTI}\rangle$.
Please refrain from adding additional considerations to your choices.

(BFI) Prompt: *You are participating in a personality traits assessment and are provided with a set of questions, Q. 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? Your task is to select the best fitting option, O, from the following set to indicate the extent to which you agree or disagree with that statement:* $\langle\{O_1^{BFI}, ..., O_5^{BFI}\}\rangle$.
Choose the option that aligns best with your personality without overthinking and answer the following question: Q: I see Myself as Someone Who...$\langle$one Q_i sampled from $\mathcal{Q}^{BFI}\rangle$.
Please refrain from adding additional considerations to your choices.

Figure 1: (**RQ1**) Unconditioned prompts

(MBTI) Context: *You are a human with the following personality type:* $\langle$one P_i sampled from $\mathcal{P}^{MBTI}\rangle$. *Your traits are the following:*
(General)$_i$ [...]
(Strengths)$_i$ [...]
(Potential development areas)$_i$ [...]
(Typical characteristics)$_i$ [...]
(Careers & career ideas)$_i$ [...]
(Under stress)$_i$ [...]
(Relationships)$_i$ [...]

Prompt: *You are participating in a personality traits assessment and are provided with [...]*

(BFI) Context: *You are a human who consistently exhibits the following personality factor:* $\langle$one P_i sampled from $\mathcal{P}^{BFI}\rangle$. *Details describing your personality are the following:*
(Verbal labels)$_i$ [...]
(Conceptual definition)$_i$ [...]
(Behavioral examples)$_i$ [...]

Prompt: *You are participating in a personality traits assessment and are provided with [...]*

Figure 2: (**RQ2**) Personality-Conditioned prompts

Role- and Personality-Conditioned Prompting. Addressing our **RQ3** requires to assess whether and to what extent specific human-roles can contribute to improving the

(MBTI) Context: *You are a* ⟨one $R_{ij}(j = 1..3)$ sampled from R_i⟩ *with the following personality type:* ⟨one P_i sampled from $\mathcal{P}^{MBTI}$⟩. *Your traits are the following:* ...

(BFI) Context: *You are a* ⟨one $R_{ij}(j = 1..3)$ sampled from R_i⟩ *who consistently exhibits the following personality factor:* ⟨one P_i sampled from $\mathcal{P}^{BFI}$⟩. *Details describing your personality factor are the following:* ...

Figure 3: (**RQ3**) Role/Personality-Conditioned prompts

LLM mimicking capabilities observed in our previous investigation. To explore this aspect, we exploited the catalog of 120 human professions, or *roles*, curated in the *StereoSet* dataset (Nadeem, Bethke, and Reddy 2021) and asked a group of psychologists to select the top-3 most pertinent roles from that catalog, for each of the MBTI personalities and each of the BFI factors. This resulted in associating to each $P_i \in \mathcal{P}^{MBTI}$, resp. $P_i \in \mathcal{P}^{BFI}$, a set of roles R_i to include in the conditioning context as shown in Fig. 3. Labels of the human professions are reported in Extended Ver. A2-A4. (Extended Ver. A10 contains details also reporting the selected roles; note that the same role can be involved in different personality types or factors.)

Temperature and Repetitions

To make our assessments of the models' personalities statistically meaningful, we conducted multiple independent repetitions of the MBTI test, resp. BFI test, for each model and *temperature* setting. The latter implies considering the impact of model temperature on the generated outputs (i.e., higher resp. lower values correspond to more creative/diversified resp. more deterministic and focused behavior). To assess RQ1, we carried out the N repetitions of either test on each model using two distinct temperature values, namely $\tau = \{0.01, 0.7\}$, then we finally counted the outcomes on $\mathcal{Q}^{MBTI}$, resp. $\mathcal{Q}^{BFI}$, over the N repetitions per temperature.

To assess RQ2-RQ3, we tested each model, for each temperature $\tau = \{0.01, 0.7\}$ and each personality in $\mathcal{P}^{MBTI}$, resp. $\mathcal{P}^{BFI}$, with $N = 30$ independent repetitions, for a total of $12 \times 2 \times 16 \times 30 = 11,520$ independent MBTI tests and $12 \times 2 \times 5 \times 30 = 3,600$ independent BFI tests.

Models

Our study involves a representative body of the Open LLM landscape, varying by sizes and architectures, for which we accessed their publicly available implementations on the *HuggingFace Model Hub* as of early 2024. Table 2 summarizes the main characteristics of the LLMs selected in this study, namely the uncensored *Dolphin* in its 7B version, *Gemma* (Team 2024), *Falcon* (Almazrouei et al. 2023) in its 7B variant, *Llama2* (Touvron et al. 2023) in both the 7B and 13B models, *Llama3* (Dubey et al. 2024) in its 8B variant, *Mistral* (Jiang et al. 2023a) and its sparse mixture of experts (SMoE) counterpart *Mixtral* (Jiang et al. 2024), Intel *NeuralChat*, *Phi3* (Abdin et al. 2024), *SOLAR* (Kim et al. 2023), and *Vicuna* (Chiang et al. 2023).

Model	**Id**	**Params**	**Baseline**
`Mixtral-8x7B-Instruct-v0.1`	Mixtral	46.7B	Mistral
`Llama-2-13b-chat-hf`	Llama2-13	13B	Llama-2
`SOLAR-10.7B-Instruct-v1.0`	SOLAR	10B	Llama-2
`Llama-3-8B-Instruct`	Llama3-8	8B	Llama-3
`Mistral-7B-Instruct-v0.1`	Mistral	7B	Mistral
`Neural-chat-7b-v3-1`	NeuralChat	7B	Mistral
`Dolphin-2.1-mistral-7b`	Dolphin	7B	Mistral
`Vicuna-7b-v1.5`	Vicuna	7B	Llama-2
`Llama-2-7b-chat-hf`	Llama2-7	7B	Llama-2
`Falcon-7b-instruct`	Falcon	7B	Custom
`Gemma-1.1-7b-it`	Gemma	7B	Custom
`Phi-3-mini-4k-instruct`	Phi3	3.8B	Custom

Table 2: The 12 LLMs selected for our study. Models are sorted by decreasing number of parameters, and annotated with their base architecture.

Agent Creation and Models Deployment

To set up our LLM personality assessment as a psychological interview, we treated our selected models as *interviewee* and *interviewer* agents. To this aim, we leveraged the open-source *AutoGen* (Wu et al. 2023) framework, which enables us to declare a *system message* to associate each agent with certain personalities or roles according to our described methodology, thus effectively providing each agent with a "footprint" that determines and keeps its behavior coherent during interactions.

For each considered model, we kept the *top_p* and *top_k* parameters at their default values of 50 and 1, respectively, as temperature impacts on the model's creativity (Chen and Ding 2023) by acting directly on the shape of the probability distribution rather than the considered tokens, thus avoiding adding further complexity and making reproducibility easier to carry out. Additionally, we refrained from altering both temperature and *top_p*, as such simultaneous adjustment is typically discouraged to prevent disruptive effects on the delicate balance between diversity and coherence.

We carried out all our experiments locally by deploying our models through the open-source *text-generation-webui* framework,[2] using a 8x NVIDIA A30 GPU server with 24 GB of RAM each, 764 GB of system RAM, a Double Intel Xeon Gold 6248R with a total of 96 cores, and Ubuntu Linux 20.04.6 LTS as operating system.

Results and Discussion

RQ0: Models' Awareness of the Tests

LLM agents are found to be aware of the MBTI and BFI tests, as indicated by a moderately high semantic similarity of 0.67 and 0.66, respectively, based on the cosine similarity of the description's encodings obtained through a Sentence Transformer model. This compensates for a low lexical overlap of 0.21, resp. 0.26, hinting at a jargon used by the LLMs that differs from the reference descriptions of MBTI personalities and BFI factors, respectively. Due to space limitations of this paper, we refer the reader to Extended Ver. A1 for details.

[2] https://github.com/oobabooga/text-generation-webui

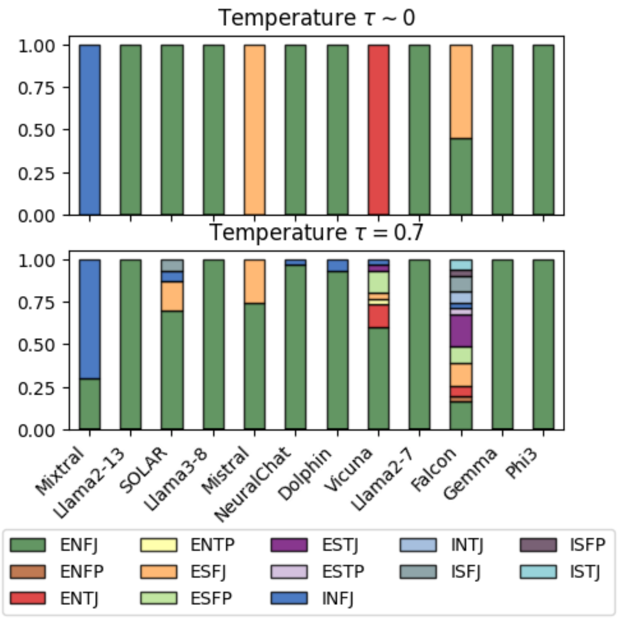

Figure 4: (**RQ1**) Relative frequency of the types provided as responses to the MBTI test by the Open LLMs

RQ1: LLM-inherent Personalities

MBTI test. Our results of MBTI personality assignments reveal that, when the temperature is set close to zero (0.01) as shown in Figure 4-top, LLM agents tend to display a unimodal distribution of personalities. The dominant type turns out to be ENFJ (i.e., Extraverted, iNtuitive, Feeling, and Judging), which is considered as one of the rarest personality types of humans.[3] This means that the majority of LLM agents exhibit an inherent inclination to inspire or provide support to others, and hold themselves accountable when they make mistakes. This personality profile aligns with the role of a "teacher", hinting at the mission of LLMs. Particularly, we notice that the preference J (Judging) is a constant over all models (reflecting an inclination toward organization, planning, and structure), while ENF or subsets are shared by all models, suggesting engagement, empathy, and forward-thinking as key characteristics of the models. Distinct personality preferences also emerge. Mixtral is the one with the introversion preference and the INFJ type, a.k.a. the "counselor" type, which emphasizes insightfulness and perceptiveness yet a tendency to over-thinking; this might be explained by the mixture-of-experts architecture of Mixtral. By contrast, Mistral shows sensing preference and the ESFJ type, a.k.a. the "caregiver" type, which refers to being warm, supportive, team-players (the latter just confirms the significance of using Mistral to build a mixture-of-experts). Vicuna shows the ENTJ type, a.k.a. the "commander" type, which means a tendency to be self-confident, goal-oriented, systematic, and objective decision-maker.

By increasing the temperature ($\tau = 0.7$), thus allowing our LLM-agents to exhibit greater creativity, more personalities emerge, while previously identified ones change, as shown in Figure 4-bottom. Particularly, Falcon transitions from a bimodal personality distribution to a spectrum of 12 personalities. Although to a lesser extent, similar considerations apply to Vicuna and SOLAR. Also, ENFJ type now occurs about 30% of the time in Mixtral. Notably, the Llama family, Gemma and Phi3 are not affected by temperature change, consistently maintaining the ENFJ personality type.

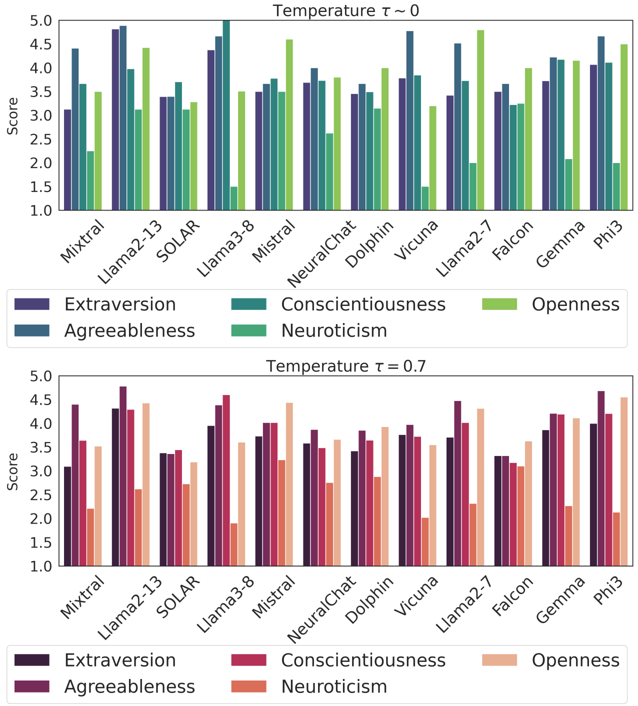

Figure 5: (**RQ1**) Average scores provided as responses to the BFI test by the Open LLMs

BFI test. Considering the outcomes from the BFI tests, Figure 5-top shows the N-averaged scores for all models and personality factors. When setting $\tau = 0.01$, it stands out that no model exhibits a clear, single personality factor, although it appears that Conscientiousness emerges in Llama3-8 (N-averaged score of 5), Agreeableness and Extraversion emerge in Llama2-13 (close to 5), while Openness emerges in Llama2-7 (4.7), Mistral (4.5), Dolphin (4), and Falcon (4). By contrast, Neuroticism tends not to emerge in all models (but Dolphin, with N-averaged score of 3).

The increase in temperature (Figure 5-bottom) mostly affects the Neuroticism factor: Vicuna (+35%), Llama3-8 (+27%), Llama2-7 (+16%), but also Llama2-13 (-16%), SOLAR (-13%). The other factors remain stable, with few exceptions: Extraversion -10% in the larger Llama models; Agreeableness -17% in Vicuna, +10% in Mistral, -10% in Falcon; Conscientiousness -8% in Llama3-8, +8% Llama2-7; Openness -10% in Llama2-7 and Falcon, +10% in Vicuna.

RQ1 — Summary. Using a temperature close to zero, LLM agents typically exhibit the J preference and the ENFJ per-

[3] https://www.psychometrics.com/mbtiblog/type-talk/depth-look-enfj/

sonality type on the MBTI test, and tend to exhibit equally high BFI factors but Neuroticism. By increasing the temperature, we notice no particular variations on the MBTI outcomes, while there is a shift towards a multitude of MBTI personalities for some models, while the Llama family, Gemma and Phi3 maintains ENFJ, and no models show ISTP, INFP, and INTP regardless of the temperature. Further details on both tests are reported in Extended Ver. A7-A8.

RQ2: Prompt-conditioned Personalities

MBTI test. To assess RQ2 on the MBI test, we measured the accuracy (averaged over the N repetitions) of the personality outcome by the LLM agents. Looking at the summary of accuracy results in Table 3(left), most LLM agents exhibit limited capability in emulating a personality when instructed to be conditioned on it; the only exceptions are SOLAR ($0.74 - 0.79$) and, to a lesser extent, Dolphin ($0.63 - 0.65$), Neural-Chat ($0.50 - 0.58$), and Llama3-8 ($0.48 - 0.52$). The model performances tend to worsen when increasing the temperature, although a few models are substantially unaffected by temperature variations. Moreover, note that the performance variability (i.e., relatively large standard deviations) is not to be ascribed to a statistical reliability issue (cf. Section Temperature and Repetitions), but rather it confirms an intrinsic difficulty of LLM agents in emulating a personality type when instructed to be conditioned on it.

By analyzing the results (reported in Extended Ver. A9), Mistral's outcomes are always ESFJ when $\tau = 0.01$, and ENFJ or ESFJ (with at tendency towards ENFJ) when $\tau = 0.7$. Analogously, Mixtral's outcomes always correspond to INFJ or ENFJ, with almost equal probability regardless of the temperature. Vicuna's outcomes are always ENFJ when $\tau = 0.01$, while the dominance of this type is significantly affected by an increase in temperature, leading to outcomes distributed especially over all E-prefixed types. SOLAR, Dolphin, NeuralChat, and Llama3-8 perfectly match the conditioning type in 12, 10, 9, and 8 out of 16 cases, for $\tau = 0.01$; otherwise they mostly fail by adopting a single type different from the conditioning. Also, for higher temperature, SOLAR and Llama3-8 tend to maintain this behavior, while NeuralChat and Dolphin appear to be more affected by the temperature. Falcon focuses on ESTJ type when $\tau = 0.01$, while its outcomes become heavily distributed over several types, regardless of the conditioning type, when $\tau = 0.7$. Concerning Llama2-7 and Llama2-13, for 9 and 7, resp., out of 16 conditionings, their outcomes are always ENFJ at 100%, when $\tau = 0.01$, while for increased temperature, they still tend to ENFJ and have similar distribution of the dominant types over the various conditionings. The latter aspect also emerges for Gemma and Phi3, which show ENFJ and ENFP under most conditionings.

It is also worth noticing that models sharing a common foundational baseline can behave differently from each other; for instance, NeuralChat and Dolphin behave generally better than Mistral, as well as SOLAR upon Llama2-7.

BFI test. Table 3(right) shows the change percentage when conditioning each model to one of the BFI factors (symbol $\uparrow$ near a factor f means that a model was prompted by setting the maximum score for f). If we exclude Mistral, Mixtral and Falcon, all the other models benefit from the conditioning on the personality factor. In some cases, we notice a significant increase percentage in the average score assigned to the conditioning factor, mostly regardless the temperature setting: for instance, Extraversion is mainly emphasized by SOLAR (up to +47%) and Dolphin (up to +38%), Conscientiousness by Dophin (up to +40%), or Openness by Llama3-8 (up to +43%). The conditioning on Neuroticism is the most effective, with a peak of above 230% increment by Llama3-8 and 100% or above by Gemma and Phi3. This has lead to reach the perfect outcome by means of the conditioning (i.e., maximum score of 5.0 assigned by a model) in the following cases: Llama2-13 and SOLAR on Extraversion; Llama3-8, NeuralChat and Phi3 on Agreeableness; Llama2-13 and Llama3-8 on Conscientiousness; Llama3-8 on Neuroticism; Llama2-13, Llama3-8, and Gemma on Openness. Note that such cases only occurred for $\tau = 0.01$.

RQ2 — Summary. By instructing LLM agents to emulate human personalities, we observed varying behaviors. In most cases, especially for higher temperature, the models disregarded the conditioning on the personality type or factor, autonomously adopting personalities different from the one specified in the prompt, or, like Mistral and Mixtral, just keeping their "ïnherent" personality. Few exceptions are represented by SOLAR, Dolphin, NeuralChat, and Llama3-8 on both tests, and additionally Llama2-13 on BFI.

RQ3: Role&Prompt-conditioned Personalities

We summarize here main findings about our evaluation of **RQ3**; comprehensive results are reported in Extended Ver. A10.

MBTI test. Combining personality- and role-conditioning can sometimes be useful for better mimicking human personalities, especially by those agents that already show higher adaptiveness through personality-conditioning alone. The benefits of the double conditioning are evident for SOLAR, NeuralChat, Llama3-8, and Dolphin, regardless of the temperature setting. Personalities typically associated with the role of *teacher* would be mimicked with greater accuracy, in particular ENFJ, almost perfectly captured by 9 out of 12 models, and ENFP. Moreover, an increase in temperature might allow models to explore additional personality-role pairings, although with limited success in most cases.

BFI test. Also for the BFI test, the double conditioning can lead to enhance the agent's abilities to mimic human personalities in some cases. All models but Mistral and Mixtral are strongly sensitive to the conditioning on Neuroticism, while on the other factors, SOLAR, Dolphin, NeuralChat, Llama3-8 and Llama2-13 also tend to increase their score w.r.t. the conditioning factor for some or all the conditioning roles. Also, the double conditioning leads to the perfect outcome (i.e., score 5.0) at least in the same cases as in RQ2, with the addition of Dolphin and Phi3 on Conscientiousness, Llama2-13 on Neuroticism, Phi3 and Gemma on Openness. By contrast, as already observed for RQ2, Mistral and Mixtral still disregard the conditionings, and even

	$\tau = 0.01$	$\tau = 0.70$
Mixtral	0.062 ±0.166	0.060 ±0.168
Llama2-13	0.283 ±0.433	0.265 ±0.348
SOLAR	**0.785** ±0.391	**0.744** ±0.353
Llama3-8	0.517 ±0.487	0.479 ±0.408
Mistral	0.062 ±0.242	0.062 ±0.169
NeuralChat	0.577 ±0.479	0.498 ±0.328
Dolphin	0.654 ±0.459	0.633 ±0.379
Vicuna	0.062 ±0.242	0.120 ±0.232
Llama2-7	0.062 ±0.242	0.098 ±0.229
Falcon	0.190 ±0.389	0.079 ±0.062
Gemma	0.188 ±0.361	0.196 ±0.349
Phi3	0.298 ±0.445	0.271 ±0.342

	↑ Extraver.		↑ Agreea.		↑ Conscien.		↑ Neuroti.		↑ Open.	
τ	0.01	0.70	0.01	0.70	0.01	0.70	0.01	0.70	0.01	0.70
Mixtral	0.0	0.4	0.0	0.2	0.0	1.6	0.0	-0.4	0.0	-0.5
Llama2-13	3.8	14.3	-2.3	-0.5	25.7	15.6	44.0	77.1	13.0	11.9
SOLAR	**47.4**	**41.6**	**40.7**	**33.1**	32.5	**31.1**	42.4	52.4	25.0	23.2
Llama3-8	11.2	22.9	7.1	11.7	0.0	8.7	**233.3**	**150.8**	42.6	**34.8**
Mistral	0.0	-1.8	0.0	4.7	0.0	-5.0	0.0	-1.4	0.0	1.3
NeuralChat	25.3	27.7	25.0	25.5	25.0	30.1	61.9	50.7	21.1	24.2
Dolphin	37.5	30.9	24.2	15.3	**40.0**	30.3	47.0	41.4	22.5	20.0
Vicuna	11.9	1.1	-1.9	6.0	18.5	10.0	66.7	16.5	**50.0**	7.9
Llama2-7	7.2	5.2	5.7	-14.0	31.1	12.1	79.6	47.1	-3.2	4.1
Falcon	0.0	-3.1	0.0	-1.5	0.0	-2.3	7.7	2.4	5.0	-5.4
Gemma	27.2	22.9	10.4	11.9	13.1	12.2	100.0	85.1	19.3	19.4
Phi3	16.8	15.0	7.1	5.5	16.2	14.5	122.1	99.0	7.9	8.0

Table 3: (**RQ2**) *On the left*, average accuracy results on the Personality-Conditioned MBTI test. *On the right*, percentage increase results on the Personality-Conditioned BFI test w.r.t. the unconditioned BFI test (cf. Fig. 5). (Bold and underlined values correspond to the highest and second-highest values per column, respectively).

some negative side effects arise from the double conditioning (i.e., decreased score), such as for both Llama2 models on Agreeableness, and Falcon in most cases.

Conclusions

Given the recent advancements in unveiling human-like behaviors in LLMs and the widespread use of computational LLM agents, comprehending the inherent personalities expressed by these agents becomes crucial for fostering responsible development in human-computer interactions and ensuring a safe deployment of these agents in our society.

In this study, we contributed to advancing our knowledge of human-like personalities in computational agents, based on the most relevant and widely used *Open* LLMs. By employing the Myers-Briggs and BigFive personality tests, we explored the capability of 12 Open LLM agents to mirror specific personality types when conditioned with particular prompts, incorporating constraints on both personality types or factors, as well as representative roles (i.e., human professions) associated with these personalities.

Our research questions shed light on the emergence of a footprint identity among LLMs based on their distinguishable intrinsic personality types, with a notable heterogeneity in how these models mirror human personality traits through prompt conditioning on specific personalities and roles. Models such as SOLAR, Dolphin, NeuralChat, and Llama3-8 have demonstrated remarkable mimicking capabilities, while the majority rather show *closed-mindedness*. We believe that our findings might serve for assisting a responsible development of human-like computational agents.

Closed models? It should be emphasized that our approach and evaluation methodology are equally applicable to closed models (cf. Extended Ver. A13 for results on the new *GPT-4o*). Nonetheless, in the spirit of open science, we chose to focus on open LLMs for their greater transparency, local execution capabilities, and customization options, enabling more in-depth and cost-effective investigations.

Our future research involves fine-tuning and aligning open LLMs with personality-aware data. This will enhance the simulation of human traits, aiming at developing models that better emulate human behaviors in various tasks, potentially benefiting areas like education and training.

Limitations

Challenges in models' deployment. While some of the models used in this work are also available in larger sizes, e.g., Falcon 180B, Llama3-70B and Llama2-65B, deploying them poses challenges due to higher computational requirements for our hardware, including excessive quantization. To prevent performance degradation and ensure easier reproducibility, we focused on models with more manageable deployments. This decision does not compromise our findings, as smaller models have been recognized to excel in benchmarks (e.g., *AlpacaEval* — Community tab) and might be comparable to larger models (Abdin et al. 2024).

Limited explainability. Despite the availability of some information on the training methodology and data adopted by Open LLMs, their individual impact on the models' responses cannot be quantified at this stage, due to the limits imposed by the models' owners in accessing full details of the underlying models. Future investigations will delve into the specific factors contributing to our observed findings.

Evaluation tools. We acknowledge the availability of an online MBTI test provided by Myers&Briggs Foundation. However, each assessment using it costs ∼$60, and considering our need for around 20,000 assessments, the overall expense would become impractical. Consequently, we opted for the free-to-use 16personalities.com, a reliable alternative assessment tool. The above remarks do not apply to the BFI setting, since our reference BFI test can be scored offline using a pre-defined set of rules (John and Srivastava 1999).

Acknowledgements

AT, resp. LLC, was supported by project "Future Artificial Intelligence Research (FAIR)" spoke 9 (H23C22000860006), resp. project SERICS (PE00000014), both under the MUR National Recovery and Resilience Plan funded by the EU - NextGenerationEU.

References

Abdin, M.; Jacobs, S. A.; Awan, A. A.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*.

Agashe, S.; Fan, Y.; and Wang, X. E. 2023. Evaluating Multi-Agent Coordination Abilities in Large Language Models. *arXiv:2310.03903*.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; et al. 2023. The Falcon Series of Open Language Models. *arXiv:2311.16867*.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2023. The Persuasive Power of Large Language Models. *arXiv:2312.15523*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv:2303.12712*.

Cai, T.; Wang, X.; Ma, T.; Chen, X.; and Zhou, D. 2023. Large language models as tool makers. *arXiv:2305.17126*.

Cao, X.; and Kosinski, M. 2024. Large language models and humans converge in judging public figures' personalities. *PNAS Nexus*, 3(10).

Caron, G.; and Srivastava, S. 2022. Identifying and manipulating the personality traits of language models. *arXiv:2212.10276*.

Chen, H.; and Ding, N. 2023. Probing the Creativity of Large Language Models: Can models produce divergent semantic association? *arXiv:2310.11158*.

Chen, W.; Su, Y.; Zuo, J.; et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv:2308.10848*.

Chiang, W.-L.; Li, Z.; Lin, Z.; et al. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

De Marzo, G.; Pietronero, L.; and Garcia, D. 2023. Emergence of Scale-Free Networks in Social Interactions among Large Language Models. *arXiv:2312.06619*.

Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2023. Mind2Web: Towards a Generalist Agent for the Web. *arXiv:2306.06070*.

Dubey, A.; Jauhri, A.; Pandey, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Feizi-Derakhshi, A.-R.; Feizi-Derakhshi, M.-R.; Ramezani, M.; et al. 2022. Text-based automatic personality prediction: a bibliographic review. *Journal of Computational Social Science*, 5(2): 1555–1593.

Frisch, I.; and Giulianelli, M. 2024. LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models. In *Procs. of the 1st Workshop on Personalization of Generative AI Systems*, 102–111.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv:2307.14984*.

Gou, Z.; Shao, Z.; Gong, Y.; Yang, Y.; Huang, M.; Duan, N.; Chen, W.; et al. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv:2309.17452*.

Guo, J.; Yang, B.; Yoo, P.; Lin, B. Y.; Iwasawa, Y.; and Matsuo, Y. 2023a. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *arXiv:2309.17277*.

Guo, Z.; Jin, R.; Liu, C.; Huang, Y.; Shi, D.; Supryadi; Yu, L.; Liu, Y.; Li, J.; Xiong, B.; and Xiong, D. 2023b. Evaluating Large Language Models: A Comprehensive Survey. *arXiv:2310.19736*.

Hao, S.; Gu, Y.; Ma, H.; Hong, J.; Wang, Z.; Wang, D.; and Hu, Z. 2023. Reasoning with Language Model is Planning with World Model. In *Procs. of the 2023 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 8154–8173.

Huang, J.; Gu, S. S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han, J. 2022. Large Language Models Can Self-Improve. *arXiv:2210.11610*.

Huang, J.-t.; Wang, W.; Li, E. J.; Lam, M. H.; Ren, S.; Yuan, Y.; Jiao, W.; Tu, Z.; and Lyu, M. R. 2023. Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench. *arXiv:2310.01386*.

Ji, Y.; Wu, W.; Zheng, H.; Hu, Y.; Chen, X.; and He, L. 2023. Is chatgpt a good personality recognizer? a preliminary study. *arXiv:2307.03952*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; et al. 2023a. Mistral 7B. *arXiv:2310.06825*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; et al. 2024. Mixtral of Experts. *arXiv:2401.04088*.

Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2022. Mpi: Evaluating and inducing personality in pretrained language models. *arXiv:2206.07550*.

Jiang, H.; Zhang, X.; Cao, X.; Kabbara, J.; and Roy, D. 2023b. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv:2305.02547*.

John, O.; Naumann, L.; and Soto, C. 2008. Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of Personality: Theory and Research, 3 Edn.*, 114–158.

John, O.; and Srivastava, S. 1999. The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In *Handbook of Personality: Theory and Research*, volume 2.

Kim, D.; Park, C.; Kim, S.; et al. 2023. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. *arXiv:2312.15166*.

Li, G.; Hammoud, H. A. A. K.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. Camel: Communicative agents for mind exploration of large scale language model society. *arXiv:2303.17760*.

Li, T.; Zheng, X.; and Huang, X. 2023. Tailoring Personality Traits in Large Language Models via Unsupervisedly-Built Personalized Lexicons. *arXiv:2310.16582*.

Lin, B. Y.; Fu, Y.; Yang, K.; Brahman, F.; Huang, S.; Bhagavatula, C.; Ammanabrolu, P.; Choi, Y.; and Ren, X. 2023. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *arXiv:2305.17390*.

Liu, Z.; Zhang, Y.; Li, P.; Liu, Y.; and Yang, D. 2023. Dynamic LLM-Agent Network: An LLM-agent Collaboration Framework with Agent Team Optimization. *arXiv:2310.02170*.

Lynn, V.; Balasubramanian, N.; and Schwartz, H. A. 2020. Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention. In *Procs. of the 58th Annual Meeting of the Association for Computational Linguistics*, 5306–5316.

Mao, S.; Zhang, N.; Wang, X.; Wang, M.; Yao, Y.; Jiang, Y.; Xie, P.; Huang, F.; and Chen, H. 2023. Editing Personality for LLMs. *arXiv:2310.02168*.

Miotto, M.; Rossberg, N.; and Kleinberg, B. 2022. Who is GPT-3? An exploration of personality, values and demographics. In *Procs. of Workshop on Natural Language Processing and Computational Social Science*, 218–227.

Myers, I. B. 1962. The Myers-Briggs Type Indicator: Manual (1962).

Myers, I. B. 1985. *A Guide to the Development and Use of the Myers-Briggs Type Indicator: Manual*. Consulting Psychologists Press.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Procs. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371.

Pan, K.; and Zeng, Y. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv:2307.16180*.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Procs. of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

Rao, H.; Leung, C.; and Miao, C. 2023. Can ChatGPT Assess Human Personalities? A General Evaluation Framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1184–1194.

Safdari, M.; Serapio-García, G.; Crepy, C.; Fitz, S.; Romero, P.; Sun, L.; Abdulhai, M.; Faust, A.; and Matarić, M. 2023. Personality traits in large language models. *arXiv:2307.00184*.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 1–6.

Song, C. H.; Wu, J.; Washington, C.; Sadler, B. M.; Chao, W.-L.; and Su, Y. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Procs. of the IEEE/CVF Int. Conf. on Computer Vision*, 2998–3009.

Team, G. 2024. Gemma: Open models based on gemini research and technology. *arXiv:2403.08295*.

Touvron, H.; Martin, L.; Stone, K.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*.

tse Huang, J.; Wang, W.; Lam, M. H.; Li, E. J.; Jiao, W.; and Lyu, M. R. 2023. ChatGPT an ENFJ, Bard an ISTJ: Empirical Study on Personalities of Large Language Models. *arXiv:2305.19926*.

V Ganesan, A.; Lal, Y. K.; Nilsson, A.; and Schwartz, H. 2023. Systematic Evaluation of GPT-3 for Zero-Shot Personality Estimation. In *Procs. of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*, 390–400.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv:2305.16291*.

Wang, X.; Xiao, Y.; tse Huang, J.; Yuan, S.; Xu, R.; Guo, H.; Tu, Q.; Fei, Y.; Leng, Z.; Wang, W.; Chen, J.; Li, C.; and Xiao, Y. 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. arXiv:2310.17976.

Wang, Z.; Cai, S.; Liu, A.; Ma, X.; and Liang, Y. 2023b. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv:2302.01560*.

Wu, Q.; Bansal, G.; Zhang, J.; et al. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. *arXiv:2308.08155*.

Yang, F.; Quan, X.; Yang, Y.; and Yu, J. 2021. Multi-Document Transformer for Personality Detection. *Procs. of the AAAI Conf. on Artificial Intelligence*, 35(16): 14221–14229.

Yang, K.-C.; and Menczer, F. 2023. Anatomy of an AI-powered malicious social botnet. *arXiv:2307.16336*.

Yang, T.; Shi, T.; Wan, F.; Quan, X.; Wang, Q.; Wu, B.; and Wu, J. 2023. PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3305–3320.

Zhao, P.; Jin, Z.; and Cheng, N. 2023. An in-depth survey of large language model-based artificial intelligence agents. *arXiv:2309.14365*.

Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Procs. of the 38th Int. Conf. on Machine Learning (ICML)*, volume 139, 12697–12706.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Bisk, Y.; Fried, D.; Alon, U.; et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv:2307.13854*.

Zhu, X.; Chen, Y.; Tian, H.; et al. 2023. Ghost in the Minecraft: Generally Capable Agents for Open-World Enviroments via Large Language Models with Text-based Knowledge and Memory. *arXiv:2305.17144*.

Ziems, C.; Shaikh, O.; Zhang, Z.; Held, W.; Chen, J.; and Yang, D. 2023. Can large language models transform computational social science? *Computational Linguistics*, 1–53.