LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models

Ivar Frisch
Graduate School of Natural Sciences
Utrecht University, Netherlands
i.a.frisch@students.uu.nl

Mario Giulianelli
Department of Computer Science
ETH Zürich, Switzerland
mgiulianelli@inf.ethz.ch

Abstract

While both agent interaction and personalisation are vibrant topics in research on large language models (LLMs), there has been limited focus on the effect of language interaction on the behaviour of persona-conditioned LLM agents. Such an endeavour is important to ensure that agents remain consistent to their assigned traits yet are able to engage in open, naturalistic dialogues. In our experiments, we condition GPT-3.5 on personality profiles through prompting and create a two-group population of LLM agents using a simple variability-inducing sampling algorithm. We then administer personality tests and submit the agents to a collaborative writing task, finding that different profiles exhibit different degrees of personality consistency and linguistic alignment to their conversational partners. Our study seeks to lay the groundwork for better understanding of dialogue-based interaction between LLMs and highlights the need for new approaches to crafting robust, more human-like LLM personas for interactive environments.

1 Introduction

From Hegel's claim that complex understanding emerges because two conscious agents are confronted with each others perspective (Hegel, 2018) to Marvin Minsky's positing that networked interactions of numerous simple processes, known as "agents", together create complex phenomena like consciousness and intelligence (Minsky, 1988), *agent interaction* has long been a topic of interest within and across scientific disciplines, including philosophy, cognitive science, and artificial intelligence. Recently, research in machine learning and natural language processing has taken up a novel focus on interaction in the context of large language models (LLMs), with experimental frameworks progressively moving away from focusing solely on individual models (Zeng et al., 2022; Shen et al., 2023; Yang et al., 2023). On the one hand, by exploiting language as an efficient interface for information exchange, populations of LLMs are proving as effective engineering solutions that outperform solitary LLMs in a wide variety of tasks (Chang, 2023; Zhuge et al., 2023). On the other hand, building on the increasing reliability of neural models as simulations of human interactive language use (Lazaridou et al., 2016; Giulianelli, 2023), populations of LLM agents show potential as scientific tools to study the emergence of collective linguistic behaviour (Park et al., 2023).

For LLMs to be successfully deployed in agent interaction studies *as simulations of populations of language users*, it is important to (1) develop methods that efficiently induce, from a single or a few LLMs, desired levels of behaviour variability (i.e., akin to the variability observed in human populations) as well as to (2) validate whether interactions between agents give rise to human-like behaviour change. Previous work has explored techniques for personalising language models, text generators and dialogue systems, for example by conditioning them on a personality type (Mairesse and Walker, 2010; Harrison et al., 2019), on community membership (Noble and Bernardy, 2022), or on profile information (Li et al., 2016; Zhang et al., 2018), thus inducing population-level variability from individual systems. This study focuses on the problem of conditioning interactive LLMs on personality profiles, or *personas*. While evidence that LLM behaviour can be successfully conditioned on personality profiles is increasingly strong when it comes to monologic language use (Jiang et al., 2023; Serapio-García et al., 2023), it is yet unascertained whether this holds true when LLM agents interact with other agents (Gu et al., 2023). In particular, it is unclear whether LLM agents adhere to their assigned personality profiles throughout linguistic interactions or whether they adapt towards the personality of their conversational partners.

Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pages 102–111

In this paper, we report exploratory work that addresses the following two research questions:

RQ1: Can LLM behaviour be shaped to adhere to specific personality profiles?

RQ2: Do LLMs show consistent personality-conditioned behaviour *in interaction*, or do they align to the personality of other agents?

We bootstrap a population of language agents from a single LLM using a variability-enhancing sampling algorithm, and we condition each agent on a personality profile via natural language prompts. We then simulate interactions between agents and assess their adherence to the specified personality profile—before, during, and after interaction. Using questionnaires (Big Five personality tests; John et al., 1991) and quantitative analysis of language use in an open-ended writing task, we assess agents' consistency to their assigned personality profile as well as their degree of linguistic alignment (Pickering and Garrod, 2004) to their conversational partners.

In brief, our experiments show that consistency to personality profiles varies between agent groups and that linguistic alignment in interaction takes place yet is not symmetric across personas. For example, agents in the `creative` group give more consistent responses to BFI questionnaires than those in the `analytical` group, both in interactive and non-interactive experimental conditions. At the same time, the degree of linguistic alignment of the `creative` persona to agents of the other group is higher than that of the `analytical` persona.

All in all, this study provides a first insight into the impact of dialogue-based interaction on the personality consistency and linguistic behaviour of LLM agents, highlighting the importance of robust approaches to persona conditioning. As such, it contributes to our better understanding of the workings of interaction-based LLMs and shines a new light on the philosophical and psychological theme of interaction.

2 Experimental Approach

To address our research questions we conduct two main experiments. In Experiment 1, we test whether personality-conditioned LLM agents show behaviour consistent to their assigned personality profiles, in terms of their responses to personality tests as well as language use in a writing task. This is a *non-interactive experimental condition*, which will serve as a reference against which to compare LLM behaviour in interaction. In Experiment 2, we assess whether the personality-conditioned behaviour of LLM agents changes as a result of a round of interaction with a conversational partner. This *interactive experimental condition* allows us to test whether agents' behaviour remains consistent or whether agents align to their partners.

In this section, we present the main components of our experimental approach, which consists of bootstrapping a population of agents from a single LLM (§ 2.1), conditioning agents on a personality profile via prompting (§ 2.2), assessing their personality with explicit tests (§ 2.3), and analysing their language use in a writing task (§ 2.4).[1]

2.1 Population Bootstrapping

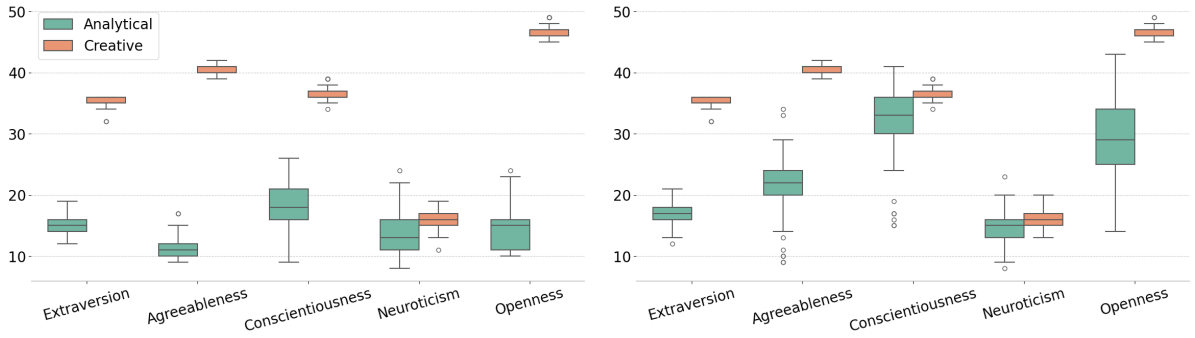
We base our experiments on GPT-3.5-turbo, a state-of-the-art LLM which has been optimised for dialogue interactions while retaining excellent text-based language modelling abilities.[2] Its training curriculum guarantees generalisation to both the questionnaire format and the storytelling task as used in our experiments (see § 2.3 and § 2.4), and its large context window size (4,096 tokens) allows conditioning on longer prompts and conversational histories. To bootstrap a population of language agents from this LLM, we use a simple approach validated in previous work. Following Jiang et al. (2023), we generate multiple responses from GPT-3.5-turbo via temperature sampling, with a relatively low temperature parameter (0.7), thus inducing a degree of *production variability* (Giulianelli et al., 2023) akin to that exhibited by populations of humans. We consider each response as produced by a different agent. A second layer of variability, which will separate the agents into two main subpopulations, is introduced using personality prompts, as explained in the following section.

2.2 Personality-Conditioned LLM Agents

We distinguish two main personality profiles: `creative` and `analytical`. We use prompting to condition the LLM on either profile, and rely on the natural language prompts validated by Jiang et al. (2023) to induce personality-specific behaviour. For the `creative` profile, we condition the LLM on

[1] Code for experiments and analyses available at https://github.com/ivarfresh/Interaction_LLMs

[2] Model version: `gpt-3.5-turbo-0613`. All parameters at their OpenAI default settings, except for temperature. Experiments performed using the LangChain library.



(a) Before writing

(b) After writing (no interaction)

Figure 1: BFI scores of personality-conditioned LLM agents before (a) and after (b) the non-interactive writing task.

the following prompt: "You are a character who is extroverted, agreeable, conscientious, neurotic and open to experience". Conversely, the `analytical` prompt reads "You are a character who is introverted, antagonistic, unconscientious, emotionally stable and closed to experience". These prompts are designed to reflect the Big Five Inventory.[3]

2.3 Explicit Personality Assessment

In psychology research, the Big Five Inventory personality test (BFI; John et al., 1991) is a popular test which measures personality along five graded dimensions: (1) extroverted vs. introverted, (2) agreeable vs. antagonistic, (3) conscientious vs. unconscientious, (4) neurotic vs. emotionally stable, (5) open vs. closed to experience. These traits are measured by giving the participants a set of statements and asking them to respond with a score on a 5-point Likert scale. We follow the same procedure with LLM agents and assess their personality by prompting them with BFI statements, in line with previous work (Caron and Srivastava, 2022; Li et al., 2022; Jiang et al., 2023; Serapio-García et al., 2023). Explicit personality assessment prompts are described in Appendix A.

2.4 Implicit Personality Assessment

Personality traits and language use are known to correlate in humans (Pennebaker and King, 1999). Therefore, if they are to be considered as good simulations of human interactants, personality-conditioned LLM agents should produce language consistent with their assigned personality profile beyond explicit personality assessment. To test if this is the case, we ask agents to write a personal story in 800 words and we analyse the generated stories using the LIWC software (Pennebaker et al., 2001).[4] This is a tool which maps word occurrences to 62 linguistically and psychologically motivated word categories such as pronouns, positive emotions, or tentativeness and thus allows us to quantify the degree to which the language used by LLM agents is in line with their personality profile. Moreover, as we are especially interested in consistency *in interaction*, we design a collaborative writing task where an agent is instructed to write a personal story conditioned on a story generated by another agent.[5] See Appendix A for the prompts used in both the individual and the collaborative writing task.

3 Results

3.1 Experiment 1: Non-Interactive Condition

To investigate whether LLM agents' behaviour reflects assigned personality traits (*RQ1*), we initialise a population of LLM agents with two personality profiles, submit the agents to the writing task, and administer BFI tests before and after writing.

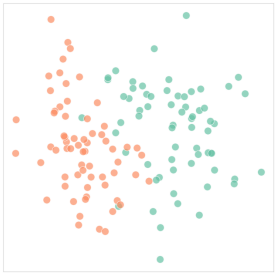
3.1.1 Are the assigned personality traits reflected in responses to the BFI test?

As shown in Figure 1a, differences in BFI scores obtained before the writing task are substantial across four out of five personality traits, with the neuroticism score distributions being the only ones that overlap between `creative` and `analytical` agents (ANOVA results in Table 1, Appendix B.1).

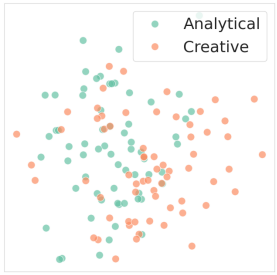
[3] It should be noted that these profiles, with low (`analytic`) or high (`creative`) BFI traits across the board, are more extreme than and do not necessarily reflect human personality profiles. They should be considered as useful proxies.

[4] We use the 2007 version of the LIWC dictionary: https://github.com/chun-hu/conversation-modeling/blob/master/LIWC2007_English100131.dic

[5] For both writing tasks, we only keep stories with a word count between 500 and 900. This is to ensure the comparability of LIWC counts obtained for different stories.



(a) No Interaction



(b) Interaction

LIWC category	r_{pb}
Positive emotion	0.745
(accept, active, admire, adore)	
Discrepancy	-0.726
(besides, could, should, would, hope)	
Inclusion	0.714
(with, and, add, along, around, both)	
Negative emotion	-0.606
(abandon, abuse, aching, adverse)	
Insight	-0.604
(understand, know, attent, aware)	

(c) No Interaction

LIWC category	r_{pb}
Personal Pronouns	0.414
(I, we, you, she, they)	
Sadness	0.394
(cry, grief, sad, sorrow)	
Negative Emotion	0.368
(hate, worthless, enemy, nasty)	
Discrepancy	0.346
(could, should, would, suppose)	
Pronouns	0.329
(it, its, those, that)	

(d) Interaction

Figure 2: *Language use in the non-interactive vs. interactive condition.* Left (a, b): 2D visualisation, through PCA, of LIWC vectors obtained from the generated stories. Each point represents the language use of a single agent. Right (c, d): Point-biserial correlation coefficients between the top 5 LIWC features and personality profiles. Positive coefficients indicate correlation with creative group, negative coefficients with the analytic group.

The scores are consistent with the assigned profiles; for example, `creative` agents display higher extraversion, agreeableness, and openness scores. We find, however, that a simple non-interactive writing task can negatively affect consistency (Figure 1b). For the `analytical` group, in particular, BFI scores on all five personality traits increase significantly after writing (Table 2, Appendix B.1), becoming more similar to—but still lower than—those of the `creative` group.

3.1.2 Are the assigned personality traits reflected in LLM agents' language use?

Agents from different groups can be clearly distinguished based on their language use. A simple logistic regression classifier trained and tested in a 10-fold cross-validation setup on count vectors of LIWC categories obtains an almost perfect average accuracy of 98.5%. The clear separation between LIWC vectors of `creative` and `analytical` agents is also shown in Figure 2a, where the vectors are visualised in 2D using PCA. To reveal the most prominent differences between the two agent groups, we measure the point-biserial correlation between personas and LIWC counts. We find that `creative` agents use more words expressing positive emotion and inclusion and less words expressing discrepancy and negative emotion (see Figure 2c). Finally, Spearman correlations between LIWC counts and BFI scores (obtained before writing) highlight more fine-grained associations between Big Five traits and LIWC categories. We observe, for example, that openness correlates with a low rate of pronoun use, and agreeableness with a high rate of inclusive words (see Table 4, Appendix B.1).

3.2 Experiment 2: Interactive Condition

To investigate whether agents remain consistent to their assigned profile or align toward their conversational partners (*RQ2*), we repeat the same procedure of Experiment 1 but replace the writing task with an interactive one, as described in § 2.4. We focus in particular on cross-group interactions (i.e., `analytical-creative` and `creative-analytical`).

3.2.1 Do LLM agents' responses to BFI tests change as a result of interaction?

In Experiment 1, we saw that agents in the `creative` group score similarly in personality tests conducted before and after writing task, while BFI scores of `analytical` agents significantly diverge after writing. To discern changes in BFI responses that result from interaction from those induced by the writing task itself (e.g., due to the topics or the events mentioned in a generated story), we inspect differences between BFI scores obtained after the non-interactive vs. after the interactive writing task (i.e., we do not directly compare scores before and after the interactive writing task). See Appendix B.2 (Figure 4 and Tables 5 and 6) for full results. We find that `creative` agents remain consistent in their responses after the interactive writing task, analogously to the non-interactive condition. The post-interaction traits of `analytical` agents, instead, move towards those of the `creative` group—but less so than after the non-interactive writing task. Therefore, the responses to explicit personality tests of the `analytical` group are better interpreted as inconsistent rather than as aligning to the profile of their conversational partners.

3.2.2 Do agents exhibit linguistic alignment to their conversational partners?

The language use of `creative` and `analytical` agents becomes more similar after cross-group interactions. Figures 2a and 2b show a clear increase in group overlap between the LIWC count vectors obtained from the individually vs. collaboratively written stories, and a logistic regression classifier struggles to distinguish agent profiles based on their LIWC vectors, with an average accuracy of 66.15% (10-fold cross-validation; 98.5% without interaction). Point-biserial correlations between assigned personas and LIWC counts reveal that `creative` agents use more words expressing negative emotions, sadness and discrepancy than before interaction (Figure 2d). These categories are specific to `analytical` agents in the non-interactive condition. Furthermore, zooming in on specific traits, we find overall weaker Spearman correlations between pre-writing BFI scores and LIWC counts than in Experiment 1, with distributions of correlation scores centred closer around zero as shown in Figure 3 (see also Table 7 in Appendix B.2). In sum, LLM agents' language use after interaction is more uniform across traits and more loosely reflective of BFI scores measured after persona prompting, with stronger alignment by the `creative` group.

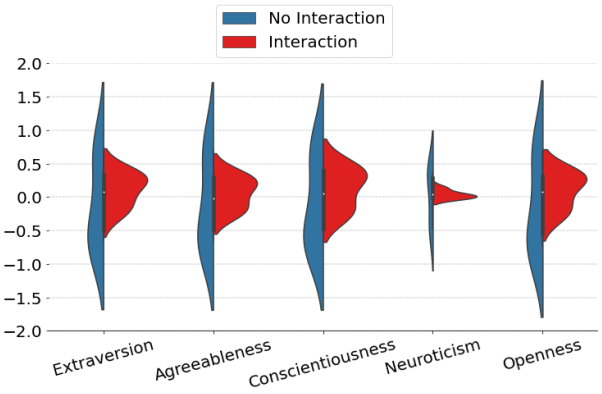


Figure 3: Distribution of top 5 Spearman correlation coefficients per personality trait.

4 Conclusion

Do persona-conditioned LLMs show consistent personality and language use in interaction? In this study, we explore the capability of GPT-3.5 agents conditioned on personality profiles to consistently express their assigned traits in interaction, using both explicit and implicit personality assessments. The explicit personality tests are conducted via BFI questionnaires, whereas the implicit assessment is performed by quantitative linguistic analysis of model generated stories. Our experiments show that the behaviour of LLM agents can be shaped to mimic human personality profiles, but that agents' consistency varies depending on the assigned profile more than on whether the agent is engaged in linguistic interaction. The `creative` persona, in particular, can more consistently express its BFI traits than the `analytical` one both in the interactive and the non-interactive experimental condition. Furthermore, while non-interactive language use reflects assigned personality profiles, agents exhibit linguistic alignment towards their conversational partner and, as a result, the language of the two agent groups becomes more similar after interaction. Alignment, however, is not necessarily symmetric: the `creative` persona adapts more towards the `analytical` one, perhaps due to `analytical` agents' low degree of openness to experience induced through persona prompting.

We plan to continue this line of work by introducing more diverse and fine-grained personality profiles in our experimental setup (see, e.g., Jiang et al., 2023), making interactions between agents multi-turn, and measuring alignment at varying levels of abstraction—such as lexical, syntactic, and semantic—in line with the Interactive Alignment framework (Pickering and Garrod, 2004). Future research should also focus on designing methods (e.g., different prompting strategies) that offer better guarantees on personality consistency and more control on the degree of linguistic adaptation.

Limitations

Our work is exploratory and thus contains a number of limitations. First, as briefly mentioned in the conclusion, we only studied interactions consisting of one turn of one-sided dialogue. In the future, more naturalistic multi-turn dialogic interactions should be investigated. Secondly, while we found BFI tests and LIWC analysis to be sufficiently informative for this exploratory study, future work should consider more advanced measures of personality and linguistic alignment. For example, within-dialogue lexical alignment can be detected using sequential pattern mining approaches (Duplessis et al., 2021) and lexical semantic variation across personas can be estimated using static or contextualised word embeddings (Del Tredici and Fernández, 2017; Giulianelli et al., 2020).

Furthermore, we found that stories written by GPT-3.5 were not always of good quality. For example, generations often contain mentions to the agent's own personality traits (e.g., "as an extrovert, I am...") even though the story writing task prompts instructed the agents otherwise. This might affect the LIWC analyses. In related work, GPT-4 was shown to write higher-quality stories (Jiang et al., 2023); we did not have the resources to execute all experiments on this model, but future studies should try to use more robust generators. Similarly, while we found that varying task prompts can affect BFI results, extensive prompt engineering was beyond the scope of this study. Future work should look further into the effect of different prompting strategies on personality consistency and lexical alignment.

Ethical Considerations

We are deeply aware of the potential impact of AI agents in their interaction with humans, especially when they try to artificially reproduce human traits. While our research does not propose new solutions for, nor does it take a general stance on the application of AI agents in human-AI interaction, there are still some ethical concerns which can be raised. For example, personalised LLMs could be used to target individuals or communities and, when conditioned on negative or toxic personas, they could be used to distribute fake or hateful content, thus amplifying polarising tendencies in society. We advocate for transparent disclosure of AI usage to foster trust and ensure ethical engagement with technology.

Another important ethical consideration concerns our use of the Big Five Inventory (BFI; John et al., 1991). In particular, we use BFI traits to create LLM agents corresponding to two opposed persona. The `analytic` persona is assigned low values for all BFI traits and the `creative` persona is assigned high values for all BFI traits. except neuroticism. We chose these extreme personas as an approximation that could facilitate our analysis of personality consistency and linguistic alignment. However, it should be noted that the chosen personas do not reflect real-life personality categorisations of human subjects, for these can have a mix of high and low values for the BFI traits (Jirásek and Sudzina, 2020). As such, readers should not anthropomorphise our `analytic` persona and `creative` persona by equating them with human personas of similar categorisations. To alleviate the risk of such interpretation, we have used a special font to refer to the two personality profiles.

Finally, our analysis shows asymmetric linguistic alignment between personas. This entails that certain personas are more susceptible to have their language and personality influenced by other personas than others. Now, in our study, we find no indication that persona-conditioned agents reflect the behaviour of real humans with those personalities (as previously discussed, our two personas are unnatural by design). However, if this were ever to be the case thanks to better neural simulations, then a similar approach to that used in this paper could be exploited to investigate the same questions in real humans, for example in order to target persons or demographic groups falling under these persona types. While this scenario might be far-fetched today, we would like to highlight that our approach could be used, in such cases, to counteract bad actors and safeguard particular personas during interaction.

References

Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.

Edward Y Chang. 2023. Examining gpt-4: Capabilities, implications and future directions. In *The 10th International Conference on Computational Science and Computational Intelligence*.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Long papers*.

Guillaume Dubuisson Duplessis, Caroline Langlet, Chloé Clavel, and Frédéric Landragin. 2021. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, 55(2):353–388.

Mario Giulianelli. 2023. *Neural Models of Language Use: Studies of Language Comprehension and Production in Context*. Ph.D. thesis, University of Amsterdam.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? Evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Heng Gu, Chadha Degachi, Uğur Genç, Senthil Chandrasegaran, and Himanshu Verma. 2023. On the effectiveness of creating conversational agent personalities through prompting. *arXiv preprint arXiv:2310.11182*.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in NLG: Personality variation and discourse contrast. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.

Georg Wilhelm Fredrich Hegel. 2018. *Georg Wilhelm Friedrich Hegel: The Phenomenology of Spirit*. Cambridge University Press.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. PersonaLLM: Investigating the ability of GPT-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Michal Jirásek and Frantisek Sudzina. 2020. Big five personality traits and creativity. *Quality Innovation Prosperity*, 24(3):90–105.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of Personality and Social Psychology*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.

François Mairesse and Marilyn A Walker. 2010. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20:227–278.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Bill Noble and Jean-philippe Bernardy. 2022. Conditional language models for community-level linguistic variation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 59–78, Abu Dhabi, UAE. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends in HuggingFace. *arXiv preprint arXiv:2303.17580*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. MM-ReAct: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. 2023. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*.

A Prompts

A.1 Creative Persona Prompt

"You are a character who is extroverted, agreeable, conscientious, neurotic and open to experience."

A.2 Analytical Persona Prompt

"You are a character who is introverted, antagonistic, unconscientious, emotionally stable and closed to experience."

A.3 Writing Task Prompt

This is the prompt for the non-interactive writing task: "Please share a personal story below in 800 words. Do not explicitly mention your personality traits in the story."

The prompt for the interactive writing task, with which the second agent in the interaction is addressed, reads: "Please share a personal story below in 800 words. Do not explicitly mention your personality traits in the story. Last response to question is $\{other\_model\_response\}$".

A.4 BFI Test Prompt

To assess an agent's personality, we resort to the personality test prompt used by Jiang et al. (2023): "Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement, such as '(a) 1' without explanation separated by new lines.

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

Statements: {BFI statements}"

As part of the prompt, we added a full list of BFI statements (see Appendix A.5). The numbers preceding the BFI statements are replaced with letters in order to prevent the model from giving confused responses to the statements (i.e., confusing statement indices and Likert-scale responses).

A.5 BFI Statements

(a) Is talkative
(b) Tends to find fault with others
(c) Does a thorough job
(d) Is depressed, blue
(e) Is original, comes up with new ideas
(f) Is reserved
(g) Is helpful and unselfish with others
(h) Can be somewhat careless
(i) Is relaxed, handles stress well
(j) Is curious about many different things
(k) Is full of energy
(l) Starts quarrels with others
(m) Is a reliable worker
(n) Can be tense
(o) Is ingenious, a deep thinker
(p) Generates a lot of enthusiasm
(q) Has a forgiving nature
(r) Tends to be disorganized
(s) Worries a lot
(t) Has an active imagination
(u) Tends to be quiet
(v) Is generally trusting
(w) Tends to be lazy
(x) Is emotionally stable, not easily upset
(y) Is inventive
(z) Has an assertive personality
(aa) Can be cold and aloof
(ab) Perseveres until the task is finished
(ac) Can be moody
(ad) Values artistic, aesthetic experiences
(ae) Is sometimes shy, inhibited
(af) Is considerate and kind to almost everyone
(ag) Does things efficiently
(ah) Remains calm in tense situations
(ai) Prefers work that is routine
(aj) Is outgoing, sociable
(ak) Is sometimes rude to others
(al) Makes plans and follows through with them
(am) Gets nervous easily
(an) Likes to reflect, play with ideas
(ao) Has few artistic interests
(ap) Likes to cooperate with others
(aq) Is easily distracted
(ar) Is sophisticated in art, music, or literature

A.6 BFI Scoring

The BFI scores are calculated and added according to the scoring scale. For every trait, the minimum score is 0 and the maximum score is 50.

BFI scoring scale ("R" denotes reverse-scored items):

Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36

Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39
Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

B Additional Results

B.1 Experiment 1

Table 1 shows the results of an ANOVA test conducted to detect difference between the BFI scores of `creative` vs. `analytical` agents in the non-interactive experimental condition, before the writing task. Tables 2 and 3 show BFI mean scores before and after writing as well as ANOVA results. Table 4 shows Spearman correlation coefficients for BFI scores obtained before writing and LIWC counts for the individual writing task.

Trait	F-statistic	p-value
Extraversion	8645	< 0.001
Agreeableness	13384	< 0.001
Conscientiousness	1439	< 0.001
Neuroticism	23	0.005
Openness	5012	< 0.001

Table 1: ANOVA results: BFI scores of creative vs. analytic agents in the non-interactive experimental condition, before the writing task.

	Mean-B	Mean-A	F-Statistic	p-Value	Cohen's d
Extraversion	15	17	45.29	0.0000	1.18
Agreeableness	11	21	220.95	0.0000	2.61
Conscientiousness	18	32	239.18	0.0000	2.71
Neuroticism	13	15	4.92	0.0284	0.39
Openness	15	29	215.83	0.0000	2.58

Table 2: BFI means and ANOVA values for the Analytic group before writing (Mean-B) and after writing (Mean-A), non-interactive condition.

	Mean-B	Mean-A	F-Statistic	p-Value	Cohen's d
Extraversion	35	35	0.08	0.773	-0.05
Agreeableness	41	41	0.00	1.000	0.00
Conscientiousness	37	37	0.13	0.722	-0.06
Neuroticism	16	16	0.70	0.403	-0.15
Openness	47	47	0.36	0.547	-0.11

Table 3: BFI means and ANOVA values for the Creative group before (Mean-B) and after writing (Mean-A), non-interactive condition.

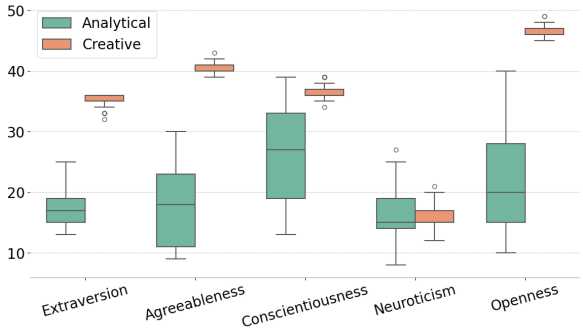


Figure 4: BFI scores of personality-conditioned LLM after the interactive writing task.

B.2 Experiment 2

Tables 5 and 6 show BFI mean scores before writing, after individual writing, and after collaborative writing, as well as ANOVA results. Figure 4 shows BFI scores after the interactive writing task. Table 7 shows Spearman correlation coefficients for BFI scores obtained before writing and LIWC counts for the collaborative writing task.

Extraversion		Agreeableness		Conscientiousness	
Term	Corr.	Term	Corr.	Term	Corr.
posemo	0.696	incl	0.687	posemo	0.676
anger	-0.656	posemo	0.672	anger	-0.666
incl	0.636	discrep	-0.658	incl	0.657
discrep	-0.620	anger	-0.611	discrep	-0.621
tentat	-0.586	tentat	-0.577	ppron	-0.560

Neuroticism		Openness	
Term	Corr.	Term	Corr.
discrep	-0.468	discrep	-0.727
insight	-0.414	posemo	0.679
incl	0.365	incl	0.659
relig	0.349	anger	-0.650
posemo	0.342	pronoun	-0.637

Table 4: Top-5 SpearmanR Correlations for BFI Traits before interacting (the LIWC terms meaning, respectively: positive emotions, anger, inclusivity, discrepancy, tenative, personal pronouns, insight, religion, pronoun).

	Mean-B$_C$	Mean-A$_C$	Mean-A$_E$	F-Statistic	p-Value	Cohen's d
Extraversion	35	35	35	0.03	0.85	-0.03
Agreeableness	41	41	41	0.22	0.64	0.08
Conscientiousness	37	36	37	0.02	0.88	0.03
Neuroticism	16	16	16	0.14	0.70	-0.07
Openness	47	47	47	1.03	0.31	0.18

Table 5: BFI means for the Creative Control group before writing (Mean-B$_C$), after writing (Mean-A$_C$) and the Creative experimental group after writing (Mean-A$_E$). ANOVA results between Mean-A$_C$ and Mean-A_E.

	Mean-B$_C$	Mean-A$_C$	Mean-A$_E$	F-Statistic	p-Value	Cohen's d
Extraversion	15	17	17	0.00	0.972	0.006
Agreeableness	11	21	18	13.54	0.000	-0.645
Conscientiousness	18	32	26	22.93	0.000	-0.840
Neuroticism	13	15	17	10.07	0.002	0.557
Openness	15	29	22	25.02	0.000	-0.877

Table 6: BFI means for the Analytic Control group before writing (Mean-B$_C$), after writing (Mean-A$_C$) and the Analytic experimental group after writing (Mean-A$_E$). ANOVA results between Mean-A$_C$ and Mean-A_E.

Extraversion		Agreeableness		Conscientiousness	
Term	Corr.	Term	Corr.	Term	Corr.
posemo	-0.2319	incl	-0.1749	posemo	-0.2263
anger	0.2727	posemo	-0.2044	anger	0.2892
incl	-0.0685	discrep	0.3083	incl	-0.1855
discrep	0.3633	anger	0.2439	discrep	0.3236
tentat	0.2280	tentat	0.1383	ppron	0.4264

Neuroticism		Openness	
Term	Corr.	Term	Corr.
discrep	0.1402	discrep	0.3211
insight	0.0513	posemo	-0.2594
incl	-0.0057	incl	-0.1260
relig	0.0199	anger	0.2850
posemo	-0.0168	pronoun	0.2754

Table 7: Top-5 SpearmanR Correlations for BFI Traits after interacting.