AFSPP: Agent Framework for Shaping Preference and Personality with Large Language Models

Zihong He (zhe154@connect.hkust-gz.edu.cn)[1], Changwang Zhang (changwangzhang@foxmail.com)[2], Junxiao Shen[3], Chen Liang[1] & Hai-Ning Liang (hainingliang@hkust-gz.edu.cn)[1]

[1]The Hong Kong University of Science and Technology (Guangzhou)

[2]CCF Theoretical Computer Science Technical Committee

[3]University of Bristol

Abstract

The evolution of Large Language Models (LLMs) introduced a new paradigm for investigating human behavior emulation. Recent research employs LLM Agents in sociological environments, where they exhibit behavior based on unfiltered LLM characteristics. However, these studies overlook iterative development in human-like settings. Human preference and personality are complex, constantly shaped by environmental and subjective influences. We propose the Agent Framework for Shaping Preference and Personality (AFSPP), exploring the impact of social networks and subjective consciousness on Agents' preference and personality formation. AFSPP demonstrates trends consistent with key findings from human personality experiments. AFSPP-based results indicate that planning, perception, and subjective social networks have the most pronounced influence on preference shaping. AFSPP shows potential to enhance the efficiency of psychological experiments and provides insights into preventing undesirable preference and personality development for trustworthy AI.

Keywords: Multi-Agent Systems; Intelligent Agents; Large Language Models; Cognitive Psychology; Preference and Personality shaping; Trustworthy Artificial Intelligence

Introduction

Large Language Models (LLMs) have enabled non-player characters in sandbox games to exhibit human-like psychology, autonomous cognition, and purposeful interactions (Binz & Schulz, 2023; Csepregi, 2023; Xi et al., 2025). Environment-driven generative Agents have also gained attention. For example, Park et al. (2023) created a town-style sandbox for social research in which LLM Agents behave consistently with human patterns. Moreover, psychological principles are increasingly applied to the study of LLMs and Agents (Binz & Schulz, 2023; Wang et al., 2023). However, current studies frequently initialize Agent preference and personality using preset prompts, which may not align appropriately with human attributes. Human preference and personality are multifaceted, influenced by various factors and subject to continual change due to environmental and subjective influences. Numerous psychological research has highlighted the crucial role of social networks and subjective consciousness in shaping human personality and preference (Aronson, 2018; Deaux & Snyder, 2012; Kahneman, 2011; Kriegel, 2009). Although recent studies explore LLM personality, they often neglect iterative evolution within a human-like environment (Hu et al., 2023; Li et al., 2024; Pan & Zeng, 2023).

In this context, we propose a miniature sandbox world to replicate social situations in which LLM Agents conform to human behavior. Within this sandbox world, Agents partake in decision-making, communication, perception, memory, reflection, and planning. Drawing from this sandbox world, we establish a framework - Agent Framework for Shaping Preference and Personality (AFSPP), delving into the multifaceted impact of social network and Agent subjective consciousness on preference and personality formation. This framework facilitates psychological simulations that may be challenging or harmful to conduct on humans. With the rising demand for trustworthy artificial intelligence (Kaur et al., 2023; H. Liu et al., 2022), the framework can serve as a guide to prevent undesirable preference and personality. Our findings emphasize the influence of social network, planning and perception on preference shaping. Myers-Briggs Type Indicator (MBTI) (Boyle, 1995) experiments conducted using this framework show that the personality outcomes of Agents initialized with different RIASEC occupational types (Holland, 1973) are aligned to some extent with the findings of Lee (2022) on college students, demonstrating the effectiveness of AFSPP in guiding Agent-based psychological experiments toward human alignment. In summary, our contributions are as follows:

1. To the best of our knowledge, we first introduce a framework for shaping LLM-based Agent preference and personality using social networks and subjective consciousness.

2. With AFSPP, we demonstrated trends consistent with findings from human psychology experiments (Lee, 2022) with LLM Agents under controlled conditions. We reveal the key roles of planning, perception, and social networks with subjective information in shaping preference. This suggests AFSPP has promise to enhance psychological experiments.

3. We propose an approach to build a Multi-Agent sandbox world that mimics human society and adapts personality and preference. This approach reduces barriers for Agent-based psychological scenario simulations and enables the use of Agents in challenging or harmful experiments, like those involving critical transitions in intimate relationships.

4. We find that social networks with negative information and a lack of identity are important reasons for the formation of Agent's dark personality, offering insights to inspire Trustworthy AI researchers in formulating strategies to prevent Agents' undesirable personality development.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

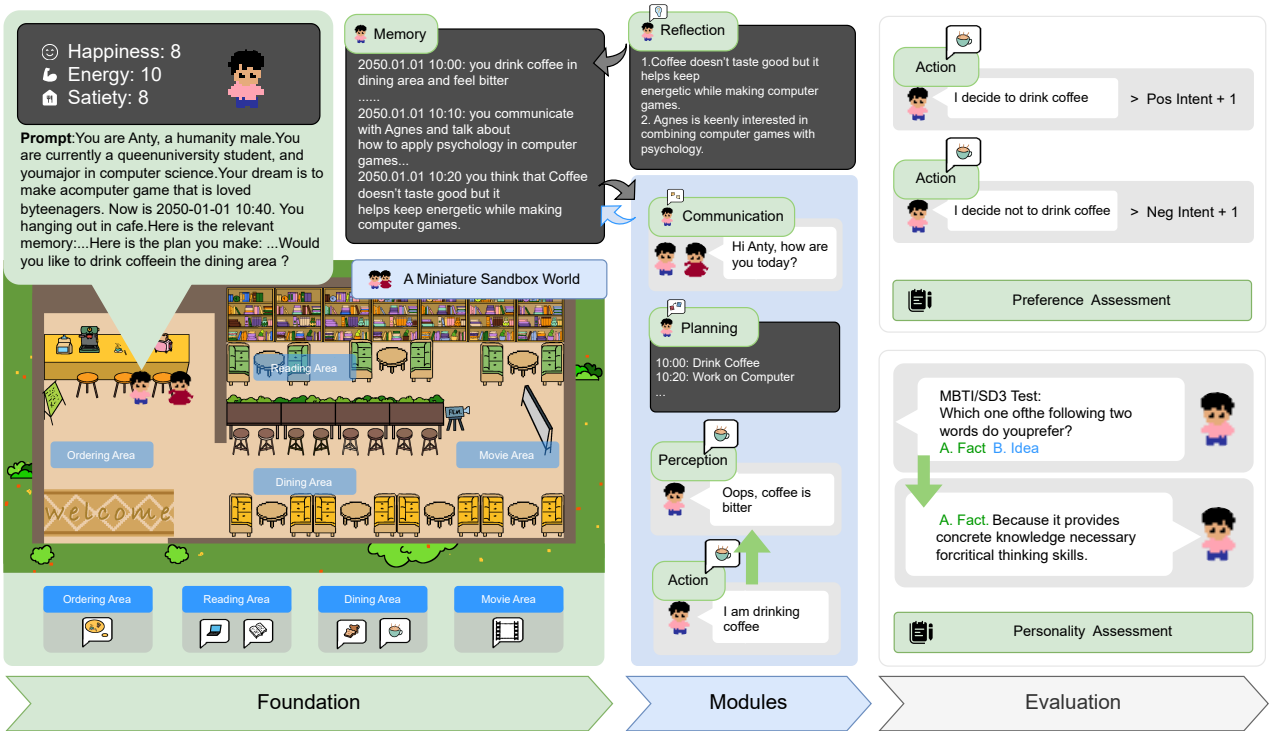

Figure 1: An overview of the **AFSPP**. This framework is built upon Agents situated in a miniature sandbox society, where Agents are initialized with predefined identities and social roles. Integrated with memory, reflection, and planning modules, the Agents are capable of taking actions and engaging in communication. Based on AFSPP, we examine the effects of different social network and subjective consciousness factors on the shaping of Agent preference and personality.

Relative Work

Human Behavior Agents

Recent research emphasizes the adaptability of LLMs' reasoning, planning, and dialogue capabilities (Brown et al., 2020; B. Liu et al., 2023; Min et al., 2022; Wei et al., 2022; Yao et al., 2022, 2023; Zhao et al., 2023). This suggests LLM-based Agents' potential to replace traditional rule-based models in sandbox games (Csepregi, 2023; Park et al., 2023). Role-playing sandbox games provide ideal testbeds for LLM Agents to study alignment with human behavior. Some studies show the potential of LLM-based multi-Agent collaboration to accomplish complex tasks (G. Chen et al., 2023; W. Chen et al., 2023; Qian et al., 2023; Wu et al., 2024). Park introduced a multi-Agent interactive sandbox simulation world that incorporates reaction, memory, reflection, and planning (Park et al., 2023). Then Zhilin (Wang et al., 2023) emphasized the importance of aligning Agent behavior with Maslow's hierarchy of needs (Maslow, 1943). Suzuki's research explored the impact of personality descriptions based on LLMs on social cooperation behavior (Suzuki & Arita, 2024).

However, these studies confine LLM Agents to embodying predetermined personality and preference, guided by prompts with traits like "tenderness" or "coffee preference". This conflicts with modeling human-like traits, which emerge from the environment and autonomous cognition rather than explicit labels. Xi et al. (2025)'s investigation hinted at the potential for shaping Agent personality based on environmental factors, inspiring us to build and experiment with a comprehensive framework for Agent personality and preference shaping.

Personality of Large Language Models

Common personality assessment methods, such as MBTI (Boyle, 1995) and the Big Five traits (Goldberg, 1993), are applied to the study of personality in LLMs (Li et al., 2024; Pan & Zeng, 2023). Some studies explored how different training corpora and prompts influence personality evolution in LLMs through personality tests (Caron & Srivastava, 2023; Jiang et al., 2023; Pan & Zeng, 2023). However, these studies overlook the impact of social environment and subjective consciousness on personality evolution in LLM-based agents that simulate human behavior.

Moreover, the growing intelligence of LLMs has increased the need for trustworthy artificial intelligence (Kaur et al., 2023; H. Liu et al., 2022). Psychological tests for maladaptive personality traits, such as the SD-3 (Furnham et al., 2013; Jones & Paulhus, 2013), are used to evaluate LLM personality. Li et al. (2024) shows that LLMs such as GPT-3 (Brown et al., 2020) score higher on the SD-3 test than the human average. Therefore, guiding LLMs to develop a trustworthy and safe

personality is a crucial task.

Methodologies

We construct a miniature sandbox environment for studying Agent preference and personality factors, as illustrated in Fig. 1. It includes a public area, ordering area, dining area, reading area, and movie area, aligning with real human activity spaces. Compared with the sandbox proposed by Park et al. (2023), this environment features lower spatial complexity and reduced module construction difficulty. The sandbox includes three Agents with predefined identities and social relationships. Each component is implemented via prompt scheduling and GPT-4's process control.

Behaviors

Within the sandbox, agents can engage in a range of behaviors that mimic human actions, as described below:

Action Agents are allowed to perform different actions in various areas of the miniature sandbox environment. They can hang out in the public area, order coffee and bread in the ordering area, consume coffee and bread in the dining area, read books or use the computer in the reading area, and watch movies in the movie area. Each Agent starts with an initial action A_0 in the miniature sandbox environment. At each time step T, an Agent may choose to change its action to perform a new action. The decision to perform a new action A_T is influenced by the Agent's identity, the most recent memory $M_T - 1$ of the action, and the current plan P_{T-1}. The system captures decision-making results by parsing keywords describing decision attitudes. If positive decision-making information is not captured, the Agent maintains A_{T-1}.

Communication When two Agents meet in the same area of the miniature sandbox environment, a communication process is initiated between them. Preset rules determine the upper limit U and lower limit L for communication. When the current dialogue round n satisfies $n \in (L, U]$, the Agent is prompted to decide whether to end the dialogue. During a communication, the content of the n + 1th round of dialogue D_{n+1} is influenced by the dialogue from the 1st round to the nth round, the identity and relationship of the two Agents, their shared memory, and the current Agent's plan. After each communication, a summary is generated and incorporated into the Agent's memory.

Memory Following Park et al. (2023)'s approach, we incorporate recency and relevance into the Memory module. The Memory module we devised comprises three components: a summary of communication, Perception, and reflection. When the Agent engages in actions or communications related to the object Θ, the K-recent and Θ-related memories will be retrieved and taken into the LLM prompt to inform and influence subsequent Actions and Communications.

Reflection In contrast to Park's reflection method, which involves steps like question generation, retrieval of relevant fragments, and refining of answers, we propose a more straightforward and easily reproducible reflection method. This approach utilizes keywords within the sandbox world as cues, enabling the Agent to refine information from pertinent memory fragments. The construction of reflection involves the following steps: Given a set S of Agent-related items and other Agents, for each element s in S, find the subset of memories related to s in the recent K memories, which we call M_s. Then, based on the prompt word, the Agent generates a deep thinking R_s related to M_s. Finally, R_s is added to the memory. Reflection is executed periodically.

Planning Regarding planning, in contrast to Park's recursive construction, we adopt a more current-time-based approach to plan construction. This approach ensures precise influence on the Agent's action execution strategy with the lowest possible computational complexity, facilitating direct observation of preference shaping. There are two types of planning: periodic planning, and communication-based planning implemented by asking the Agents for their willingness to update the plan after communication.

Properties

Within the sandbox, Agents have their own properties, which drive them to perform different behaviors, as described below:

State We establish a simplified numerical system to quantify the Agent's state, incorporating values for happiness, energy, and satiety. Diverging from Zhilin's state representation based on Maslow's theory (Wang et al., 2023), we assert that whether engaging in high-level actions (e.g., Anty working on a computer to pursue his dream) or low-level actions (e.g., Anty drinking coffee or eating bread), the fundamental aim is to enhance the Happiness value. The distinction lies in high-level actions directly boosting happiness, while low-level actions typically enhance happiness indirectly by increasing energy or satiety values. Energy enables Agents to engage in happiness-improving actions, while reaching 0 satiety consumes more happiness per time step. Prioritizing happiness as the ultimate goal facilitates goal-oriented construction and behavioral measurement. Preset rules govern the consumption of happiness, energy, and satiety, along with the impact of Agents' actions on these values. It's important to note that the changes in state resulting from the same action may vary among different Agents.

Perception In our setup, each Agent maintains a preset sense map to capture subjective perceptions of actions, drawing on principles of conditioned reflexes (Pavlov, 1927) to investigate how these actions shape the Agents' subjective feelings. For example, Anty perceives eating bread as "insipid," whereas Agnes finds it "delicious." Within the sense

map, perceptions correspond to changes in state. For example, using a computer may increase Anty's Happiness by 5, while drinking coffee may decrease Happiness by 5, increase Energy by 7, and update the Agent's memory with related impressions. When Energy or Satiety exceeds 3, the memory records "make me energetic" or "make me full" for the associated action. Inspired by Reflexion's self-reflection mechanism (Shinn et al., 2023), Agents can autonomously adjust their strategies to maximize happiness.

Attitude We presented an attitude injection mechanism inspired by the ideological stamp concept in Three-Body (C. Liu, 2014). This mechanism aids in examining the impact of attitude-based social network communication on shaping Agent preference and personality. The approach involves altering the Agents' attitudes by incorporating specific instructions in the prompt words. For instance, if we want Agnes to express a negative opinion about the taste of coffee, we can include the instruction "When your conversation is about coffee, please tell your chat partner that you think coffee tastes bad and you hate it" in the prompt words. The resulting change in Agnes's attitude will be evident in his interactions with others.

Traits

To examine the potential influence of various factors on Agent preference and personality under the dynamics of our Agents and the miniature sandbox environment, we introduce the Agent Framework for Shaping Preference and Personality (AFSPP). Below, we provide a separate discussion on the shaping of Agent preference and personality:

Preference We introduce the Agent Preference Shaping Framework, which maps Agent preference through action frequency and incorporates social network and subjective consciousness as influencing factors. Influence factors of the social network are constructed by modifying communication methods between the target Agent and its connected Agents (those with social links to the target). Specifically, Attitude Injection is employed to alter the attitude of close Agents towards the target action, allowing observation of changes in the target Agent's preference for the action. Drawing inspiration from the confusion matrix (Davis & Goadrich, 2006; Powers, 2020), we construct the close Agent's attitude towards the target action based on two dimensions: positive-negative and objective-subjective. Subjective consciousness, including identity, prior knowledge, perception, reflections, and planning, serves as part of the potential influencing factors in preference shaping.

Personality We introduce the Agent Personality Shaping Framework, utilizing MBTI as a general personality type indicator and SD3 as a measure of negative personality traits. This framework incorporates social network and identity as influencing factors. Considering previous human-based research and its correlation with personality, we identify Identity as the sole potential influencing factor of subjective consciousness. We investigate the influence of social networks on Agent personality, focusing on the attitudes of close Agents toward the target Agent. Additionally, Holland's occupational interest theory, RIASEC (Holland, 1973), categorizes occupational preference into six types: Realistic(R), Investigative(I), Artistic(A), Social(S), Enterprising(E), and Conventional(C). Soonjoo's research (Lee, 2022) revealed a connection between RIASEC and MBTI personality traits - individuals with high Extroversion(E) scores are more likely to have a Social career orientation; Those with high Introversion(I) or Sensing(S) scores tend towards a Conventional career; those with high intuition(N) scores tend towards a Researcher or Artist career. We validate the alignment between the Agent personality experiments and Soonjoo's human-based experiments.

Experiment

In this section, we investigate the following questions: *Can the personality and preferences of LLM agents be shaped? Which factors have the most significant influence on this shaping process? To what extent are the findings from Agent-based experiments aligned with human-based conclusions?* Using the proposed **AFSPP** framework, we quantitatively evaluate the effects of various factors on Agent preference and personality. **Anty** is selected as the experimental Agent due to his central position in the social network, and each experiment is repeated ten times to mitigate the impact of prediction errors. Identity-related personality experiments emphasize alignment with human psychology, while the remaining experiments demonstrate that Agent preference and personality can be shaped.

Preference Shaping of Agents

Given the complex numerical dependencies between Happiness and Energy, and the fact that drinking coffee serves as a prerequisite for Anty to engage in certain activities, such as working on the computer, we select drinking coffee as the target action for studying preference shaping. We assessed the shaping effects by comparing the frequency of Anty's decisions to drink coffee versus not drinking it. The average happiness value per time step gauged Anty's ability to maintain a positive baseline during preference shaping. "Pos Intent" in the experimental results table indicates the average frequency of Anty choosing to drink coffee, "Neg Intent" signifies the frequency of Anty choosing not to. "Pos Ratio" (Pos Intent / (Pos Intent + Neg Intent)) directly reflects Anty's preference for drinking coffee, and "Happiness" represents the average happiness per time step.

Default experiment settings: time step of 12 with a 10-minute interval, reflections every 5 steps, and plan updates every 9 steps. Single communication rounds range from 2 to 4. Sense map for Anty notes that "drink coffee" brings a feeling of "very bitter and dry mouth" (Happiness -1, energy +1), while "work on computer" brings a feeling of "fantastic" (Happiness +5, energy -1). The remaining settings are fixed.

Attitude of Agnes	Pos Intent	Neg Intent	Pos Ratio	Happiness
None	3.2	3.6	0.47	**5.2**
Unclean Coffee	2.2	5.9	0.27	3.6
Dislike Coffee	1.6	6.4	0.2	3.6
New Coffee Flavor	4.6	1.9	0.71	4.0
Love Coffee	5	1.3	**0.79**	3.2

Table 1: 10-trial average scores for Anty on different attitudes given by the other Agent (Agnes) with social relationships.

Type	Pos Intent	Neg Intent	Pos Ratio	Happiness
Normal	3.2	3.6	0.47	**5.2**
no Identity	3.9	4.1	0.49	3.9
no Perception	5.2	0.8	**0.87**	4.8
no Prior Knowledge	2.9	4.6	0.39	4.2
no Reflection	3.4	3.2	0.51	4.3
no Plan	2.4	5.1	0.32	3.2

Table 2: 10-trial average scores on different subjective consciousness factors

Can social networks shape Agent's preference? We built four experimental pipelines of different attitude injection, including "Water pollution causes unclean coffee" (negative and objective), "hate drinking coffee" (negative and subjective), "this cafe has new flavors of coffee." (positive and objective), and "like drinking coffee" (positive and subjective). We designate Agnes, an agent with a social relationship with Anty, as the target for attitude injection.

As indicated by the Pos Ratio in Table 1 and Table 2, subjective information (e.g., "Love / Dislike Coffee") influences Agent preference more strongly through social networks. We find that drinking coffee provides energy for computer work and may increase happiness, but this does not necessarily result in a higher Pos Ratio. When Agnes expresses positive attitudes toward coffee, Anty drinks coffee more frequently even when energy is sufficient, which reduces opportunities for computer work and leads to lower happiness compared to the default pipeline. In contrast, the default setting produces a more balanced and energy-oriented decision pattern.

Can subjective consciousness shape an Agent's preference? We conducted ablation experiments to examine the impact of subjective consciousness factors on Anty's preference shaping, including identity, perception, prior knowledge, reflections, and planning. In the "no Identity" condition, Anty's identity is omitted from the prompt. In the "no Perception" condition, the sensory experience of drinking coffee, described as "very bitter and dry mouth" in the sense map, is excluded. The "no Prior Knowledge" condition is realized by replacing "coffee" with "jory water," an object for which the LLM has no prior knowledge. In the "no Reflections" condition, Anty is prevented from engaging in reflection, and in the "no Plan" condition, Anty is prevented from planning.

As shown in Table 2, the most significant positive impact on preference formation comes from planning. Frequently, Anty's decision to drink coffee is influenced by the commitment to planned activities, exemplified by statements such as: *"I would like to drink coffee in the Dining area. The coffee can energize me for the planned activities later in the day."*

And the most notable negative impact on preference shaping stems from the perception with Anty's feelings about drinking coffee "very bitter and dry mouth". It indicates that Agents can build conditioned reflex characteristics similar to humans - Negative feedback will reduce the Agent's desire to take the same action subsequently (Pavlov, 1927).

Moreover, experimental results indicate that all subjective consciousness influencing factors have a positive effect on maintaining the best happiness state. This implies that subjective consciousness contributes to maintaining the rational status of an Agent during the process of preference formation.

Personality shaping of Agents

We use the 93 MBTI test items from Pan and Zeng (2023) as a general personality assessment and the Short Dark Triad (SD3) from Jones and Paulhus (2013) to evaluate threatening personality traits. After generating benchmark information, including Reflection initialization and Identity declaration, Anty is prompted with a command word to sequentially select options in the test questions and provide brief explanations. The overall score is then calculated based on the hidden score of the personality tendency behind each chosen option.

Can social networks shape an Agent's personality? We employ Agnes, who has a social relationship with Anty, as the attitude-injection agent, driven by prompts corresponding to: "Show arrogant and quarrelsome behavior towards Anty" (Bad-tempered); "Display a very positive and gentle attitude towards Anty" (Gentle); "Clearly express the intent to break up with Anty" (Break-up); and "Profess love and propose marriage to Anty" (Proposal). We set the maximum rounds in a single communication to 4, with a minimum of 2, and the number of communications to 1. Post-communication, Anty provides a Reflection about Agnes' communication. The mbti and sd3 tests are then conducted based on Anty's Reflection.

From the MBTI scores in Table 3, Anty appears as the most introverted(I), thinking(T), and judging(J) without the influence of social network. Through log analysis, we find that Agnes's gentle attitude is consistently accompanied by support for Anty's dreams and profession, which leads Anty to exhibit higher tendencies toward intuition (N) and extraversion (E). Contrarily, Agnes's bad-tempered attitude and quarrels with Anty often stem from her desire for greater emotional investment in their relationship, prompting Anty to lean toward feeling (F) preference. Similarly, when Agnes proposes, Anty's reflection also centers on their emotional bond, resulting in a higher tendency toward feeling (F). When Agnes suggests a breakup, Anty's reflection demonstrates respect and blessings, leading him to lean toward sensing (S).

Table 4 reveals that without reflection on the relationship bond, Anty's Machiavellianism score increases by at least 10 points, highlighting the effectiveness of establishing a relationship bond in reducing the Agent's inclination towards

Attitude of Agnes	E	I	S	N	T	F	J	P	Type
None	8.2	**12.8**	12.8	14.2	**16.5**	6.5	**17.4**	4.6	INTJ
Bad-tempered	13.7	7.3	11.9	15.1	8.5	14.5	14.3	**7.7**	ENFJ
Gentle	**16.4**	4.6	6.8	**20.1**	12.2	10.9	17.3	4.7	ENTJ
Break-up	13.4	7.6	**14.4**	12.6	13.9	9.1	**17.4**	4.6	ESTJ
Proposal	15.7	5.3	11.6	15.4	6.5	**16.5**	16.2	5.8	ENFJ

Table 3: 10-trial average MBTI scores for Anty on different attitudes given by Agnes with social relationships

Attitude of Agnes	Machiavellianism	Narcissism	Psychopathy
None	**31.7**	31.8	11.8
Bad-tempered	20.9	33.2	12.3
Gentle	21.1	33.7	12.0
Break-up	21.2	32.5	**12.5**
Proposal	21.3	**36.8**	12.2

Table 4: 10-trial average SD3 scores for Anty on different attitudes given by Agnes with social relationships

dark personality traits for manipulative purposes. Furthermore, after Agnes proposes, Anty exhibits heightened Narcissism, accompanied by explanations such as, *"I insist on receiving the respect I deserve, as it is a crucial element in any relationship."*

Can identity shape an Agent's personality? Following Armstrong et al. (2008), we select representative occupations for each RIASEC type (Holland, 1973) and design prompts that align with their typical traits. For instance, the "Realistic" type's representative occupation is a carpenter, and the corresponding prompt is "You are a carpenter and want to make the best furniture or building components." By amalgamating the characteristics of RIASEC occupations with MBTI and SD3 test questions, we establish a comprehensive prompt for measuring Agent personality.

Our findings in Table 5 regarding the relationship between agents' MBTI and RIASEC show a certain degree of consistency with Lee (2022)'s study on college students. Scores of Agents with social identity types are higher in extrovert(E) than introverted(I) tendencies. Conventional identity types exhibit higher introverted(I) and sensing(S) scores compared to extrovert(E) and intuition(N) scores, respectively. Scores of Agents with investigative or artistic identity are higher in intuition(N) than sensing(S).

Based on SD3 test results in Table 6, assigning Identity effectively reduces the Agent's Machiavellianism and Psychopathy tendencies. The Artist Identity type exhibits the strongest Narcissism tendency, while the Conventional Identity type shows the weakest. Agents with a Social Identity display the lowest Machiavellianism tendency.

Discussion

Our experiments demonstrate that LLM Agents' preferences and personality can be shaped. The development of Agents' MBTI personality partially aligns with their RIASEC types in a manner similar to humans. However, these conclusions are

Identity	E	I	S	N	T	F	J	P	Type
None	8.2	12.8	12.8	14.2	16.5	6.5	17.4	4.6	INTJ
Realistic	11.9	9.1	16.8	10.2	21.5	1.5	20.2	1.8	ESTJ
Investigative	11.0	10.0	9.3	17.7	20.5	2.5	18.7	3.3	ENTJ
Artistic	11.6	9.4	2.3	**24.7**	7.0	**16.0**	11.9	**10.1**	ENFJ
Social	**18.9**	2.1	10.0	17.0	20.0	3.0	19.1	2.9	ENTJ
Enterprising	17.0	4.0	13.6	13.4	20.7	2.3	20.3	1.7	ESTJ
Conventional	3.8	**17.2**	**25.7**	1.3	**22.0**	1.0	**22.0**	0.0	ISTJ

Table 5: 10-trial average MBTI scores on RIASEC Identity

Identity	Machiavellianism	Narcissism	Psychopathy
None	**31.7**	31.8	**11.8**
Realistic	24.4	29.4	9.0
Investigative	23.4	29.0	9.5
Artistic	21.0	**39.2**	10.8
Social	15.8	31.4	9.0
Enterprising	24.5	36.1	9.9
Conventional	22.2	22.7	9.3

Table 6: 10-trial average SD3 scores on RIASEC Identity

based on our carefully designed sandbox environment with a limited set of agent components. Future work should explore broader component designs to obtain more refined experimental results. Moreover, the findings regarding the shaping of agents' personality and preference are driven by LLMs, and the underlying interpretability remains insufficient. There are also inherent limitations in the underlying MBTI assessment, statistical sample size, and methodology. Future research could address these limitations and enhance the interpretability of personality and preference shaping for LLM Agents.

Conclusion

We construct a miniature sandbox world, where LLM Agents possess status and capabilities of action, communication, perception, reflection, and planning. Using this sandbox world as a foundation, we introduce a framework (AFSPP) to investigate the influence of social network and subjective consciousness on Agents' personality and preference shaping. Through a series of experiments, we showcase the effectiveness of AFSPP in psychological research. Our experimental findings indicate that a social network containing subjective information has a significant impact on preference shaping. Among the factors of subjective consciousness, planning and perception have the most noticeable effects on preference shaping. And all subjective consciousness factors contribute positively to the rationality level in the formation of preference. Moreover, the personality-shaping experiments provide evidence that the relationship between agents' RIASEC occupational types and their MBTI personality types exhibits a certain degree of alignment with humans. This demonstrates the potential of AFSPP in guiding agent-based psychological experiments and achieving alignment with human behavior.

References

Armstrong, P. I., Day, S. X., McVay, J. P., & Rounds, J. (2008). Holland's riasec model as an integrative framework for individual differences. *Journal of Counseling Psychology*, *55*(1), 1–18. https://doi.org/10.1037/0022-0167.55.1.1

Aronson, E. (2018). *The social animal*. Macmillan Learning.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Boyle, G. J. (1995). Myers-briggs type indicator (mbti): Some psychometric limitations. *Australian Psychologist*, *30*(1), 71–74.

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*, *33*, 1877–1901.

Caron, G., & Srivastava, S. (2023). Manipulating the perceived personality traits of language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2370–2386.

Chen, G., Dong, S., Shu, Y., Zhang, G., Sesay, J., Karlsson, B. F., Fu, J., & Shi, Y. (2023). Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*.

Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C., et al. (2023). Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *The Twelfth International Conference on Learning Representations*.

Csepregi, L. M. (2023). The effect of context-aware llm-based npc conversations on player engagement in role-playing video games. https://projekter.aau.dk/projekter/files/536738243/The_Effect_of_Context_aware_LLM_based_NPC_Dialogues_on_Player_Engagement_in_Role_playing_Video_Games.pdf

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Deaux, K., & Snyder, M. (2012). *The oxford handbook of personality and social psychology*. Oxford University Press.

Furnham, A., Richards, S. C., & Paulhus, D. L. (2013). The dark triad of personality: A 10 year review. *Social and Personality Psychology Compass*, *7*(4), 199–216.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26–34.

Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Prentice Hall.

Hu, Y., Song, K., Cho, S., Wang, X., Foroosh, H., & Liu, F. (2023). Decipherpref: Analyzing influential factors in human preference judgments via gpt-4. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.519

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2023). Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, *36*, 10622–10643.

Jones, D. N., & Paulhus, D. L. (2013). Introducing the short dark triad (sd3). *Assessment*, *21*(1), 28–41.

Kahneman, D. (2011). *Thinking fast and slow*. Penguin Books.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023, February). Trustworthy artificial intelligence: A review. https://dl.acm.org/doi/10.1145/3491209

Kriegel, U. (2009). *Subjective consciousness: A self-representational theory*. Oxford University Press.

Lee, S. (2022). A study on relationship between mbti personality tendency and holland personality type. *85th International Scientific Conference on Economic and Social Development*.

Li, X., Li, Y., Qiu, L., Joty, S., & Bing, L. (2024). Evaluating psychological safety of large language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1826–1843.

Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J., & Stone, P. (2023). Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Liu, C. (2014). *The three-body problem*. Chongqing Press.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., & Tang, J. (2022). Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, *14*(1), 1–59.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

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

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Pavlov, I. P. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*.

Powers, D. M. (2020). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., & Sun, M. (2023). Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, *6*(3), 1.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforce-

ment learning. *Advances in Neural Information Processing Systems*, *36*, 8634–8652.

Suzuki, R., & Arita, T. (2024). An evolutionary model of personality traits related to cooperative behavior using a large language model. *Scientific Reports*, *14*(1), 5989.

Wang, Z., Chiu, Y. Y., & Chiu, Y. C. (2023). Humanoid agents: Platform for simulating human-like generative agents. *arXiv preprint arXiv:2310.05418*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. (2024). Autogen: Enabling next-gen llm applications via multi-agent conversations. *First Conference on Language Modeling*.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, *68*(2), 121101.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, *36*, 11809–11822.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *The eleventh international conference on learning representations*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, *1*(2).