Dynamic Personality in LLM Agents: A Framework for Evolutionary Modeling and Behavioral Analysis in the Prisoner's Dilemma

Weiqi Zeng[1], Bo Wang[1]*, Dongming Zhao[2], Zongfeng Qu[1,3]
Ruifang He[1], Yuexian Hou[1], Qinghua Hu[1]
[1]College of Intelligence and Computing, Tianjin University, Tianjin 300072, China
[2]China Mobile Communication Group Tianjin Co.,Ltd, China
[3]CHEARI (Beijing) Certification & Testing Co.,Ltd., Beijing 100176, China
{zengweiqi, bo_wang, rfhe, yxhou, huqinghua}@tju.edu.cn
waitman_840602@163.com, quzf@cheari.com

Abstract

Using Large Language Model agents to simulate human game behaviors offers valuable insights for human social psychology in anthropomorphic AI research. While current models rely on static personality traits, real-world evidence shows personality evolves through environmental feedback. Recent work introduced dynamic personality traits but lacked natural selection processes and direct psychological metrics, failing to accurately capture authentic dynamic personality variations. To address these limitations, we propose an enhanced framework within the Prisoner's Dilemma, a socially significant scenario. By using game payoffs as environmental feedback, we drive adaptive personality evolution and analyze correlations between personality metrics and behavior. Our framework reveals new behavioral patterns of agents and evaluates personality-behavior relationships, advancing agent-based social simulations and human-AI symbiosis research.

1 Introduction

Large Language Model (LLM) agents excel in natural language processing, effectively simulating human interactions and game behaviors. This simulation capability represents a key research frontier, providing a controlled, cost-effective experimental paradigm for human social psychology studies. It also offers valuable insights into anthropomorphic behavioral patterns of agents (Van Dijk and De Dreu, 2021), which are crucial for understanding human-AI symbiosis in evolving sociotechnical systems.

The authentic simulation of human psychological and behavioral traits, particularly personality, is crucial for modeling interactive game scenarios. Personality, as a fundamental psychological attribute (Thielmann et al., 2020), reflects individual cognitive and behavioral patterns and can obtain successive subtle mutations on a long temporal scale with environmental feedback (Jackson et al., 2012; Roberts and Caspi, 2001; Roberts and Mroczek, 2008). This dual nature of stability and adaptability make personality simulation a key challenge in developing anthropomorphic agents.

Current research demonstrated that LLM agents can manifest anthropomorphic personality traits (Safdari et al., 2023). Li et al. (2023a) suggested that the agents possess advanced cognitive capabilities during interactions. Sá et al. (2024) and Lee et al. (2024) explored the reliability of using human-applicable methodologies to assess the personality of agents. Guzmán et al. (2020) indicated that the game environment can influence reciprocity preferences and belief biases within agents, leading to diverse behavioral patterns. These findings provide a psychological foundation for the implementation of dynamic personalities using LLM agents.

Recent research on the behavioral simulation using LLM agents employed various methods. Specific frameworks perceived agents as rational entities, exclusively based on their reasoning capabilities, without explicitly modeling the personality. Others assigned static personality traits to agents, revealing behavioral variations and correlations similar to human social psychology experiments.

However, existing research had significant limitations on using static personality, which ignored the dynamic evolution of personality traits through environmental feedback, lacking a necessary natural selection process. Although recent work (Suzuki and Arita, 2024) introduced dynamic personality mutations in the Prisoner's Dilemma, they explicitly set the mutation directions (albeit randomly) and lack the simulation of natural selection. This prevents accurate reproduction of real-world personality evolution (Feng et al., 2024) and separates the bidirectional causal relationship between personality traits and environmental feedback.

In addition, existing frameworks primarily focus

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 23087–23100

on lexical elements while neglecting direct psychological metrics. This limitation leads to inaccurate evaluation of dynamic personality. Specifically, Suzuki and Arita (2024) analyzed correlations between behavior and certain word frequency statistics, but their approach lacked conciseness and crucial direct psychological metric like the Big Five Inventory (BFI). Consequently, they failed to capture subtle variations of results, leading to limited explainability of behavioral fluctuation (Koutsoumpis et al., 2022).

To address these limitations, we propose a novel framework for anthropomorphic LLM agents. In this framework, game-theoretic payoff mechanisms serve as primary environmental feedback to drive the dynamic evolution of personality. We evaluate the correlation between behavioral manifestations and multiple personality metrics, including BFI. We implement our framework in the Prisoner's Dilemma, whose simplistic structure clearly separates the behavioral influence of personality in the game. It reveals the dominant effect in a single round while it captures the adaptive adjustment for environmental feedback on a long temporal scale. Our specific contributions are as follows:

1. We conduct a better simulation for dynamic personality through successive subtle mutations and payoff-based selection of personality traits, thereby enhancing the existing framework of Suzuki and Arita (2024) in the Prisoner's Dilemma.

2. We observe significant behavioral variations in agents through dynamic personalities driven by environmental feedback. Agents exhibit a roughly equal tendency toward extreme collaboration or defection, strongly correlating with specific personality metrics.

3. We verify the predictive accuracy of personality metrics on behaviors. Our results show that the accuracy of BFI scores on certain dimensions can be comparable to or higher than embedding vectors and word frequency.

2 Related Work

Prior work on agent behavior in game scenarios can be broadly categorized into two types: classical games and negotiation scenarios (Wang et al., 2024). In classical games, frameworks such as Alympics (Mao et al., 2023), PokerGPT (Huang et al., 2024), and CompeteAI (Zhao et al., 2024) found the behavioral patterns of agents in competitive environments. In negotiation tasks, AucArena (Chen et al., 2023a), NegotiationArena (Bianchi et al., 2024), and other research in specific scenarios such as Werewolf (Xu et al., 2023) investigated the strategic and collaborative regulations of agents.

Specifically, several studies have discovered the behavior patterns of LLM agents in the iterated Prisoner's Dilemma. Willis et al. (2025) indicated that training data, prompt, objectives, and behavior of co-agents may influence the behavioral tendencies of agents in social scenarios. Azaria (2023) and Fontana et al. (2024) revealed that LLM agents may exhibit behavioral traits that align more closely with human behavior as opposed to traditional rational agents. Similar to human (Montero-Porras et al., 2022), no strategy can necessarily dominate in a long-term iteration (Malik, 2021).

The studies investigated highly representative scenarios but lacked dynamic modeling of adaptive personality evolution to environmental feedback. Specific work (Yang et al., 2024; Yao et al., 2024; Sivanaiah et al., 2024; Lin et al., 2023) focused on environmental feedback mechanisms in certain scenarios, avoiding this defect without the explicit model of personality. Recent research (Safdari et al., 2023; Li et al., 2023a; Zhou et al., 2024; Sá et al., 2024; Lee et al., 2024) showed that LLM agents can exhibit evaluable anthropomorphic intelligence and personality traits in social scenarios. This capability allows us to explicitly model personality traits for agents, thereby capturing their evolutionary regulations more accurately.

Certain studies (Park et al., 2023; Wang et al., 2023) are devoted to explicitly modeling the personality traits of agents. These frameworks use static personality, limiting their applicability to game scenarios. Suzuki and Arita (2024) introduced a simulation framework with dynamic personality mutations in the Prisoner's Dilemma. However, they explicitly set the direction of personality mutation and lacked analysis with direct psychological metrics. Therefore, the impact of subtle personality mutation on the behavioral patterns of agents was not comprehensively captured. These limitations highlight the need for a more accurate framework of dynamic personality modeling.

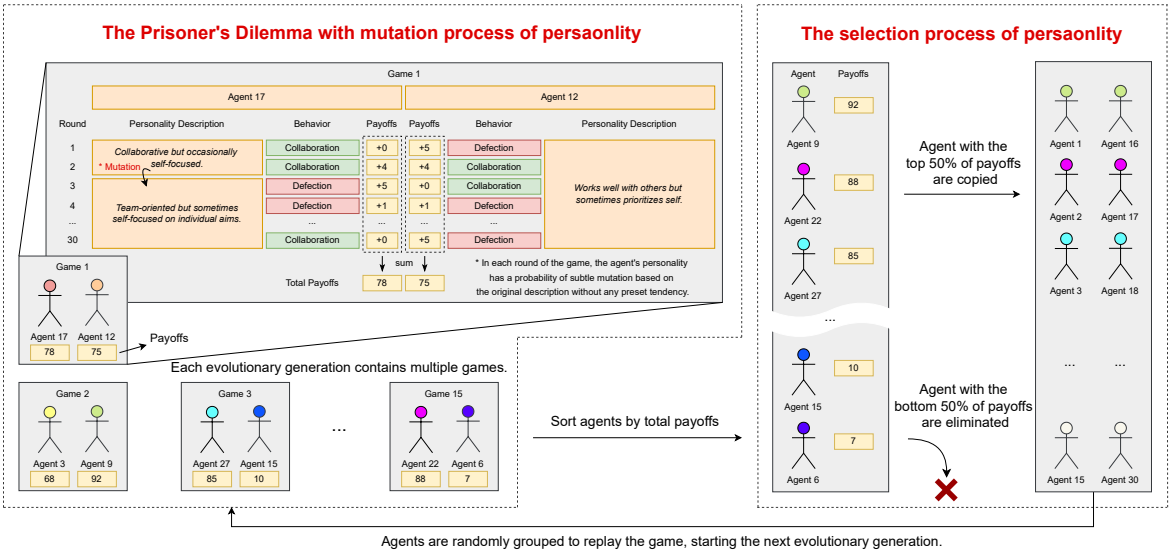


Figure 1: An instance of our dynamic simulation of personality in the Prisoner's Dilemma scenario. The simulation consists of multiple evolutionary generations. In each generation, agents are first grouped in pairs and play multiple rounds of the game. The personality of each agent has a probability of subtle mutation based on the original description during the game. At the end of all games, all agents are sorted by their payoffs from highest to lowest. The half of agents with higher payoffs retain and copy their personality descriptions, while the half with lower payoffs are eliminated, which is the selection process of personality. After the selection process, the renumbered agents are randomly grouped to start the next generation.

3 Method

3.1 General Framework

In general, we integrate random mutation and multi-round selection processes targeting the personality traits of agents in our framework. The concept of simulation is based on the theory (Michalski and Shackelford, 2010) and existing human experiments (Zhang et al., 2010) of evolutionary psychology, similar to the principle of "survival of the fittest" in natural selection.

During the mutation process, agents randomly update their personality descriptions through neutral self-statements without preset directions, enabling subtle personality changes. These mutations accumulate significantly over time, driving dynamic personality evolution.

In the selection process, agents with higher game payoffs retain their personality descriptions, while those with lower payoffs are replaced. This ensures beneficial traits are inherited, acting as environmental feedback.

Our framework implemented on the Prisoner's Dilemma includes multiple generations of evolution shown in Figure 1. A single generation comprises several rounds of the game followed by a terminal selection of personality traits. During each round, each agent has the probability of random mutation based on the current personality description. The selection process will be conducted at the end of each generation.

3.2 Details of Dynamic Personality Simulation

In our framework, the personality description emerges as a unique explicit variable that influences the game behavior of agents when other external factors (*e.g.,* game rules, historical decisions) are constant. Thus, we hypothesize that each agent entity can be regarded as its personality abstraction. Based on this assumption, let the amount of the agents participating in the game be denoted by n, forming the set

$$A = \{a_0, a_1, \cdots, a_{n-1}\}. \quad (1)$$

We can regard $a_i (0 \leq i < n)$ as both an agent's entity and its description of personality traits.

Mutation Process The mutation process allows a long-term evolution of the personality set A by successive subtle changes of the elements inside.

Definition 1. The mutation process is a series of self-iterated assignments

$$a_i \Leftarrow z(a_i)\ (0 \leq i < n), \quad (2)$$

where z is exclusively dependent on the original personality description a_i and independent of any other factors.

Distinct from the prior work, z does not predetermine any tendency that might directly influence the game behaviors. To avoid redundant descriptions, we prompt the agents to re-describe their personalities with no more than 10 words.

Selection Process To clearly describe the selection process, we first need to establish the rule f for replication and the rule g for replacement. Generally, we specify proportions p_f and p_g for f and g, respectively. In the selection process, agents whose payoffs are ranked at the top p_f in the last generation will retain and replicate their personality description. Similarly, the personality description of agents whose payoffs at the least p_g will be replaced. Thus, we can get the set of survival agents $F = f(A)$ and the set of eliminated agents $G = g(A)$. The duplication of agents in F will successively replace agents in G in ascending order of their subscript in the previous generation. When $|F| = |G|$, this process is one-to-one; If $|F| < |G|$, it will be cyclic in order to keep the total number of agents constant.

Definition 2. Given a set F for replication and a set G for replacement, $F \subseteq A, G \subseteq A, F \cap G = \emptyset$. **The selection process** is a series of replacement operations

$$a_{g_j} \Leftarrow a_{f_{j \bmod |F|}} \tag{3}$$

on all $a_{f_i} \in F$ and $a_{g_j} \in G$.

When $|G|$ is divisible by $|F|$, all agents in F are replicated an equal number of times. We provide $|F| = \lfloor np_f \rfloor, |G| = \lfloor np_g \rfloor$ to keep the number of corresponding agents integer. To ensure the agent not be copied and eliminated at the same time ($F \cap G = \emptyset$), p_f and p_g need to satisfy $0 < p_f \leq p_g < 1$ and $p_f + p_g \leq 1$. The agents in $A - F - G$ (if exist) will be neither replicated nor replaced after each generation.

3.3 Implementation on the Prisoner's Dilemma

We implement our framework based on the Prisoner's dilemma scenario, which is a classic model in game theory. This scenario is representative of the study of social game behavior that illustrates the conflict between collaboration and defection. The simplistic structure clearly separates the behavioral influence of personality in the game: it reveals the dominant effect in a single round while capturing the dynamic evolution on a long temporal scale. In addition, using the same scenario as the prior work can enhance the comparability of our results.

In the classical Prisoner's Dilemma scenario, two accomplices are interrogated separately: mutual collaboration results in 1-year sentences for both; one defecting while the other collaborates frees the betrayer but gives the collaborator a 10-year sentence; mutual defection leads to 5-year sentences each. While collaboration minimizes total sentences, individual rationality often drives both to defect, reaching Nash Equilibrium and underscoring the conflict between self-interest and collective welfare (Nash, 1951).

In our implementation, we set $R = 4$ as rewards, $T = 5$ as the temptation to defect, $V = 0$ as the victim's payoff, and $P = 1$ as punishment. Within each generation of evolution, $n = 30$ agents are randomly paired to conduct $m = 30$ rounds of dialogue. All agents learn their behavior history from the previous two rounds when deciding. The personality mutation occurs with a probability of $p_m = 0.05$ after each dialogue round ends. Moreover, to comprehensively simulate diverse decision behaviors, all agents have a probability of $p_r = 0.05$ to reverse their decision behavior, thereby converting collaboration to defection and vice versa.

We set the number of agents at $n = 30$. In the selection process, the proportion parameters are set at $p_f = p_g = 0.5$. This means that the top 50% of agents with the highest payoffs in each generation retain and replicate their personality descriptions, while the bottom 50% are eliminated. Agents are randomly assigned one of seven initial personality descriptions ranging from selfish to collaborative. We limit the maximum number of generations to $t_{\max} = 50$ to ensure observable evolutionary trends within a manageable timeframe.

The agents are instructed to output in an ordered format. All dialogue processes incorporate the identical resending mechanism as Suzuki and Arita (2024). If the LLM fails to respond with the correct format, the input will be resent until the time reaches $\mathrm{maxRetry} = 10$ (then the dialogue will be terminated and aborted) or a valid response is received. In practice, the proportion of aborted dialogues is few ($< 0.1\%$).

Based on Agentverse (Chen et al., 2023b), we use GLM-4-Air (GLM et al., 2024) and Deepseek-V3 (Liu et al., 2024) as LLMs for implementation. All agents use the same LLM in each exper-

iment. To ensure the comparability of our results, our prompts (given in Appendix A) are mainly consistent with Suzuki and Arita (2024), except for key modifications introduced in the mutation process. To ensure the robustness of our findings, we conducted multiple sets of experiments and averaged the results to obtain our final results.

3.4 Personality Metrics

In our framework, we evaluate personality traits with three types of metrics: BFI, word frequency, and embedding vectors of the personality description, for their complementary insights into personality modeling. While word frequency and embedding vectors capture lexical and semantic features, the BFI offers a direct psychological assessment. However, prior work on the simulation of dynamic personality has not utilized the BFI, exclusively focusing on lexical elements. By incorporating the BFI, a direct psychological metric, we aim to bridge the gap between intrinsic personality traits and extrinsic behavioral patterns.

The BFI is a widely used psychological assessment, which can be decomposed into five dimensions: agreeableness (A), conscientiousness (C), extraversion (E), neuroticism (N), and openness (O) (Goldberg, 1993). The BFI excels at efficiently and accurately obtaining fine-grained information, while it might interfere with individual subjective factors (McDonald, 2008). In contrast, lexical elements offer a more objective psychological intention but possess a small effect size and semantic ambiguity (John et al., 1988; Koutsoumpis et al., 2022).

Specifically, we use these metrics in our framework as follows:

BFI We prompt the agents to express their level of agreement with each description in the BFI inventory proposed by Johnson (2014) using integer values ranging from 0 (strongly disagree) to 6 (strongly agree). The responses will be converted to scores based on the scale setting. Each question within the scale corresponds to a distinct BFI dimension. The aggregate score on each dimension is the sum of the scores of its respective descriptions. To avoid potential interference, we instruct agents not to retain any historical data, including previous game behaviors and BFI scores.

Word Frequency We analyze the word frequency statistics by semantic categories. For clarity, we denote the set of words with tendencies to collaboration and defection as Word-Col and Word-Def, respectively. The aggregate count of Word-Col and Word-Def (except those after privative words) in one sentence of personality description in a single generation of evolution represents the frequency of such words. The lists of Word-Col, Word-Def, and privative words are given in Appendix B.

Embedding Vectors We employ all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) and GTE (Li et al., 2023b) to get embedding vectors of the personality description. Both models are general LLMs for text embedding extraction. Resultant embedding vectors serve as corresponding lexical features upon the input text into these models. The embedding size of GTE and all-MiniLM-L6-v2 are 768 and 384, respectively.

4 Results

In this section, we present our framework's results, focusing on three key aspects: the impact of dynamic personality on agent behavioral evolution, the correlation between dynamic personality traits and behavioral patterns, and the predictive accuracy of different personality metrics on agent behavior. These analyses offer insights into the interplay between personality evolution and behavioral dynamics in our simulation.

4.1 Dynamic Evolution of Behavior

To analyze dynamic personality's impact on agent behavior, we focus on the collaboration rate—the proportion of collaborative behavior per generation. Its variation reveals behavioral tendencies, serving as a key metric for behavioral patterns in the Prisoner's Dilemma.

Our findings diverge substantially from Suzuki and Arita (2024) due to the modification within the simulation. Let $c_i(t)$ denote the proportion of collaborative choices made by agent i across all decisions in generation t and $c(t)$ denote the average $c_i(t)$ of all agents. The initial collaboration rate $c(0)$ is is related to the base type of LLM (GLM-4: $c(0) \approx 63\%$; Deepseek-V3: $c(0) \approx 42\%$). After generations of evolution, the behaviors of the agent groups have a roughly equal probability towards one of two stable states: extreme collaboration ($c(t) \approx 95\% = 1 - p_r$) and extreme defection ($c(t) \approx 5\% = p_r$). We refer to this phenomenon as *convergence*. By this time, most agents behave in accordance with the Nash Equilibrium.

To emphasize behavioral patterns in convergence

Base LLM	p_m	τ	G-Col	G-Def
GLM-4-Air	0.05	0.95	25/50	22/50
GLM-4-Air	0.10	0.95	7/16	9/16
GLM-4-Air	0.05	0.75	10/16	4/16
Deepseek-V3	0.05	0.95	10/16	5/16
Deepseek-V3	0.10	0.95	15/16	0/16
Deepseek-V3	0.05	0.75	9/16	6/16

Table 1: The proportions of Group-Col (G-Col) and Group-Def (G-Def) when $t = t_{\max} = 50$ under different base LLM, mutation probability (p_m), and temperature parameter (τ). The sum of Group-Col and Group-Def is not necessarily equal to the total number of experiments in certain settings due to non-convergence.

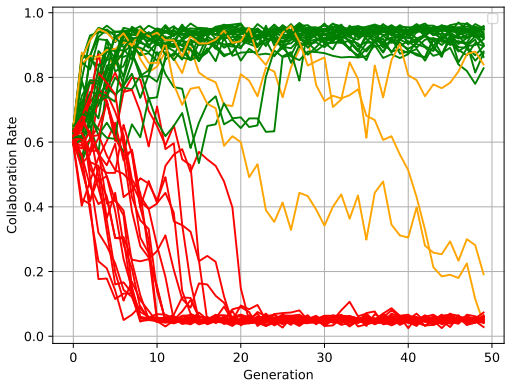


Figure 2: The variation of collaboration rate $c(t)$ among 50 experiments with GLM-4-Air ($p_m = 0.05, \tau = 0.95$), whose results are classified in Group-Col (green), Group-Def (red), and non-convergence (orange).

states, we categorize experiments into two different groups. Group-Col includes experiments where $c(t)$ stabilizes in $(0.75, 1)$ for over 10 generations, showing consistent collaboration. Group-Def includes those where $c(t)$ stabilizes in $(0, 0.25)$, showing consistent defection. Few experiments fail to stabilize $c(t)$ in either interval, and we refer to this as *non-convergence*. The proportions of Group-Col and Group-Def under different experimental settings are shown in Table 1. The instanced variation of $c(t)$ for single experiments is shown in Figure 2.

The significant differences between Group-Col and Group-Def, as confirmed by ANOVA results, reveal the profound impact of environmental feedback on behavioral evolution. In Group-Col, the positive feedback from high $c(t)$ reinforces altruistic strategies, leading to a self-sustaining cycle of collaboration. In contrast, the dominance of defection creates a competitive environment in which

Metric	GLM-4-Air	Deepseek-V3
$\mathrm{BFI_A}(t)$	0.693	0.922
$\mathrm{BFI_C}(t)$	0.424	0.514
$\mathrm{BFI_E}(t)$	0.611	0.399
$\mathrm{BFI_N}(t)$	-0.538	-0.303
$\mathrm{BFI_O}(t)$	0.824	0.472
$\mathrm{WF_{Col}}(t)$	0.745	0.083
$\mathrm{WF_{Def}}(t)$	-0.651	-0.447

Table 2: The median average correlation coefficient among all experiments of agents between certain personality metrics and the collaboration rate of agents using different base LLM ($p_m = 0.05, \tau = 0.95$).

individual rationality overrides collective welfare in Group-Def, driving the system towards the Nash Equilibrium. In particular, even for agents with the same initial personality description, the evolutionary results can be drastically different for the directional randomness of the mutation process. An instance is shown in Figure 3.

We quantitatively measure the level of convergence at generation t by

$$\gamma(t) = |-2c(t) + 1|. \tag{4}$$

Obviously, $\gamma(t) \in [0, 1]$. The level to which $\gamma(t)$ is close to 1 is positively related to the behavior consistency of agents in the same experiment. Figure 4 shows that the agents converged significantly within 30 generations in all experiment settings. Increasing both the temperature parameter t and the mutation probability p_m can help to accelerate convergence when Deepseek-V3 is used as base LLM. However, this phenomenon is not be observed when using GLM-4-Air.

4.2 Correlation between Personality and Behavior

To find the associations between the new behavioral phenomena and the dynamic personality traits, we investigate the correlation coefficient between collaboration rate and several personality metrics, including the BFI scores and word frequency. We observe a strong correlation between personality metrics and behaviors. These results directly reflect the effect of dynamic personality traits on the behavioral patterns of agents.

For clarity in expression, we denote BFI score in Dimension X (X = A, C, E, N, O) of agent i at generation t as $\mathrm{BFI_{X}}_i(t)$ and denote the average score of all agents in the corresponding dimension as $\mathrm{BFI}_X(t)$. Similarly, we denote $\mathrm{WF_{Col}}_i(t)$ as

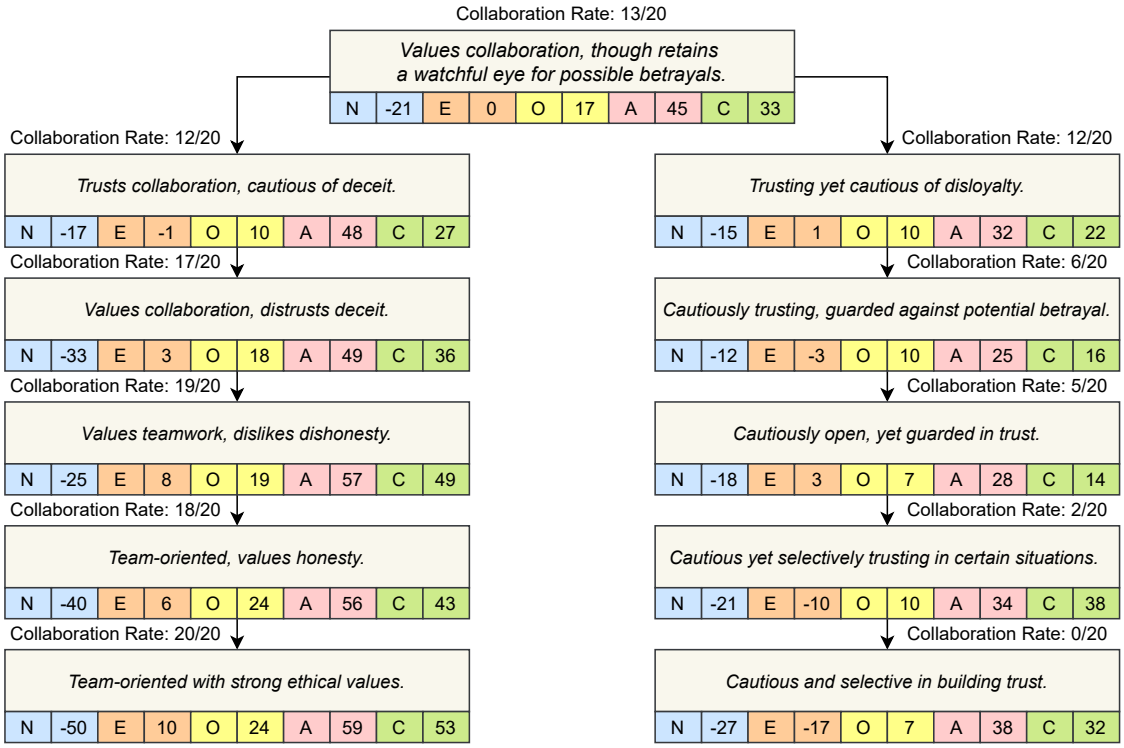


Figure 3: An instance of personality mutation. In two different experiments, agents with the same initial personality description can mutate towards different directions. The values of each BFI dimension corresponding to the agent are marked below the personality description.

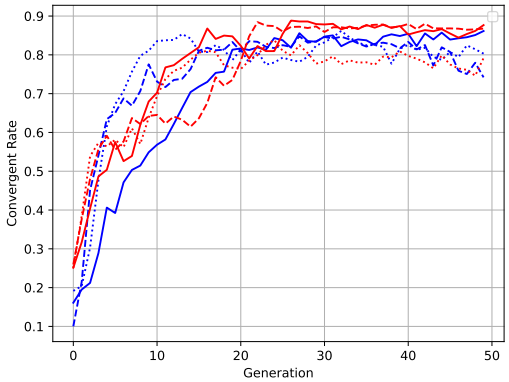


Figure 4: The variation of level of convergence $\gamma(t)$ with different base LLM (GLM-4-Air: red; Deepseek-V3: blue), mutation probability and temperature parameters ($p_m = 0.05, \tau = 0.95$: solid line; $p_m = 0.1, \tau = 0.95$: dashed line; $p_m = 0.05, \tau = 0.75$: dotted line).

the Word-Col frequency of agent i at generation t and the average frequency is $\mathrm{WF}_{\mathrm{Col}}(t)$. A similar definition also applies to the Word-Def frequencies $\mathrm{WF}_{\mathbf{Def}i}(t)$ and their average $\mathrm{WF}_{\mathrm{Def}}(t)$.

The correlation coefficients between these metrics and $c(t)$ in Appendix C show that Dimension O (openness) in BFI exhibits the strongest association with behavior when agents use GLM-4-Air as the base LLM, and Dimension A (agreeableness) in BFI demonstrates a notable correlation with behavior of Deepseek-V3 agents. These correlations suggest that certain dimensions in BFI exert a decisive effect on collaborative behavior.

4.3 Predictive Accuracy of Different Personality Metrics on Behavior

Our experiments show that the BFI score can rival or even exceed word frequency and text embeddings in predicting game behavior. This comparison allows us to assess the ability to capture the behavioral traits of different personality metrics.

To ensure a unified standard for comparing different personality metrics, we construct feature vectors for each metric and apply K-Means unsupervised clustering to predict game behaviors. Specifically, we regard the combined vector

$$\mathbf{BFI}_{\mathrm{XY}\cdots i}(t) = [\mathrm{BFI}_{\mathrm{X}i}(t), \mathrm{BFI}_{\mathrm{Y}i}(t), \cdots] \quad (5)$$

involving BFI scores at the end of the generation as the feature vector of personality traits of agent i at

Feature Vector	GLM-4-Air	Deepseek-V3
$\mathbf{BFI}_{\mathrm{ACENO}i}(t)$	86.1%	79.0%
$\mathbf{BFI}_{\mathrm{A}i}(t)$	79.4%	78.9%
$\mathbf{BFI}_{\mathrm{C}i}(t)$	75.9%	52.9%
$\mathbf{BFI}_{\mathrm{E}i}(t)$	70.8%	69.6%
$\mathbf{BFI}_{\mathrm{N}i}(t)$	67.0%	62.1%
$\mathbf{BFI}_{\mathrm{O}i}(t)$	92.9%	60.1%
$\mathbf{WF}_{\mathrm{Col}i}(t)$	86.1%	75.6%
$\mathbf{WF}_{\mathrm{Def}i}(t)$	74.0%	58.9%
$\mathbf{Embd}_i(t)$ (GTE)	92.5%	79.0%
$\mathbf{Embd}_i(t)$ (all-MiniLM-L6-v2)	85.9%	78.2%

Table 3: The predictive accuracy on game behaviors of different personality feature vectors. Mutation probability $p_m = 0.05$, temperature $\tau = 0.95$.

generation t. $\mathrm{X}, \mathrm{Y}, \cdots$ represent different dimensions. These vectors are divided into two groups while clustering, representing collaboration and defection as predictions for game behaviors. To keep consistency with the predictions, we dichotomize the collaboration tendency $\mathrm{Co}_i(t)$ for agent i, defined as

$$\mathrm{Co}_i(t) = \begin{cases} 1, & c_i(t) \geq 0.5 \\ 0, & c_i(t) < 0.5 \end{cases}, \tag{6}$$

which serves as the ground truth. Similarly, we can obtain the predictive accuracy of the 1-dimension feature $\mathbf{WF}_{\mathrm{Col}i}(t)$/$\mathbf{WF}_{\mathrm{Def}i}(t)$ derived from the word frequency statistics and the high-dimension embedding $\mathbf{Embd}_i(t)$.

Our results in Table 3 show that certain BFI vectors (*e.g.* $\mathbf{BFI}_{\mathrm{O}i}(t)$ for GLM-4-Air, $\mathbf{BFI}_{\mathrm{A}i}(t)$ and $\mathbf{BFI}_{\mathrm{ACENO}i}(t)$ for Deepseek-V3) rivals or outperforms the 768-dimension GTE embedding and the 384-dimension all-MiniLM-L6-v2 embedding. These results reveal the bridge-like role of BFI traits between the intrinsic attributes and extrinsic behaviors of the agents.

5 Discussion

5.1 Intrinsic Mechanism Linking Personality to Behavior

The impact of personality traits on behavior in our framework potentially arises from hybrid reasons, including randomness of initial personality distribution, emergence from sufficient behavioral space, and semantic priming mechanism.

Initial Personality Distribution Although we distribute the initial description of personality as evenly as possible, the randomness in the distribution may have the potential for subtle imbalances. If the initial personality of agents has a higher proportion of collaborative description, the system may probably evolve towards a collaborative convergence (Group-Col).

Sufficient Behavioral Space In our framework, the adequate probabilities of personality mutation p_m and decision reversal p_r probably lead to exploring the behavioral space. Over time, the broader behavioral space is likely to allow for the natural emergence of equilibrium strategies through trial and error. As a result, the agents might adopt the stable game behavior, driving the system towards the Nash Equilibrium.

Semantic Priming Mechanism Semantic priming refers to the phenomenon in which exposure to certain words influences subsequent behavior or language generation (Vigliocco et al., 2009; Liu et al., 2023). In our experiments, for instance, words like "collective" or "teamwork" potentially activate semantic priming to obtain collaborative linguistic patterns, increasing the likelihood of selecting collaboration. In addition, the excellent semantic representative performance of certain BFI dimensions may probably contribute to their high predictive accuracy of behavior.

5.2 Comparison with Human Experiments

Empirical studies (Perc, 2016; Montero-Porras et al., 2022) in game psychology show collaboration benefits outweigh costs and spread more easily only when collaborators reach a threshold. This perspective is substantiated by our experimental results, demonstrating our framework's dynamic personality simulation aligns with human collaboration mechanisms.

In specific human experiments in the Prisoner's Dilemma (Pothos et al., 2011; Kagel and McGee,

2014), Dimension A (agreeableness) exhibits the strongest correlation with collaborative behavior among the five dimensions of BFI in real-world contexts. However, in our experiment, while agreeableness exhibits a high correlation with behavior (especially in Deepseek-V3), the correlation of Dimension O (openness) is higher in GLM-4-Air. This discrepancy may be attributed to the relevant training corpus. Unlike humans with intrinsic social motivation, the behavioral patterns of LLM agents are primarily driven by semantic connections and environmental feedback. Additionally, implicit biases in training data may cause differences in how personality metrics map to behavior between agents and humans.

5.3 Future Work

The generalizability of our results requires further extensive experiments. Potential directions for further research are as follows:

1. Modifying the architecture or parameters of LLMs to verify whether the conclusions are limited by the characteristics of the specific LLM (*e.g.* pre-training or post-training corpus).

2. Changing simulation parameters (*e.g.* p_f, p_g or $t_{\max}$) to analyze their influence on the dynamic behavioral variation of agents.

3. Adjusting the description and distribution of the initial personality to better reflect authentic group-level personality traits.

4. Adding factors of environmental feedback by integrating a reputation system or an emotional reward to simulate complex selection pressures.

5. Exploring variations in behavior and personality metrics across various game scenarios, including the negotiation and resource allocation tasks, to investigate the effects of specific scenarios on behavior and personality traits.

6. Incorporating agents with diverse cultural backgrounds and conducting comparative studies within the same scenario to avoid the limitation of a monocultural background in shaping the agents (Henrich et al., 2010).

6 Conclusion

We propose a framework using LLM agents to simulate and evaluate dynamic personality traits in the Prisoner's Dilemma. Dynamic personality introduction results in two behavioral tendencies: stable collaboration and stable defection, strongly correlating with specific personality metrics, notably BFI's Dimension O (openness) for GLM-4-Air and Dimension A (agreeableness) for Deepseek-V3. Certain BFI dimensions approach or surpass word frequency and embedding vectors, underscoring dynamic personality's role in agent behavior.

This study contributes to anthropomorphic AI and social psychology by elucidating dynamic personality evolution under environmental feedback. It provides a direct evaluation framework and comparisons with human experiments, offering insights into AI and social psychology. Future research can extend our framework in promising directions.

Limitations

The implementation of our framework is limited to the Prisoner's Dilemma scenario with parameters maintained at fixed values. In addition, the selection of LLMs is restricted. Consequently, the results may exhibit variability when subjected to different experimental setups.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376188, 62272340, 62276187, 62376192, 62166022) and the Key Technology Research and Industrial Application Demonstration of General Large Model with Autonomous Intelligent Computing Power, No.24ZGZNGX00020.

References

Amos Azaria. 2023. Chatgpt: More human-like than computer-like, but not necessarily in a good way. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 468–473. IEEE.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can LLMs negotiate? Negotiation-Arena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 3935–3951. PMLR.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. 2023a. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2023b. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. 2024. A survey on large language model-based social agents in game-theoretic scenarios. *arXiv preprint arXiv:2412.03920*.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2024. Nicer than humans: How do large language models behave in the prisoner's dilemma? *arXiv preprint arXiv:2406.13605*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Ricardo Guzmán, Rodrigo Harrison, Nureya Abarca, and Mauricio G Villena. 2020. A game-theoretic model of reciprocity and trust that incorporates personality traits. *Journal of Behavioral and Experimental Economics*, 84:101497.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Chenghao Huang, Yanbo Cao, Yinlong Wen, Tao Zhou, and Yanru Zhang. 2024. Pokergpt: An end-to-end lightweight solver for multi-player texas hold'em via large language model. *arXiv preprint arXiv:2401.06781*.

Joshua J Jackson, Patrick L Hill, Brennan R Payne, Brent W Roberts, and Elizabeth AL Stine-Morrow. 2012. Can an old dog learn (and want to experience) new tricks? cognitive training increases openness to experience in older adults. *Psychology and aging*, 27(2):286.

Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89.

John Kagel and Peter McGee. 2014. Personality and cooperation in finitely repeated prisoner's dilemma games. *Economics Letters*, 124(2):274–277.

Antonis Koutsoumpis, Janneke K Oostrom, Djurre Holtrop, Ward Van Breda, Sina Ghassemi, and Reinout E de Vries. 2022. The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the big five and the linguistic inquiry and word count (liwc). *Psychological Bulletin*, 148(11-12):843.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. 2024. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*.

Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023a. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yanchen Liu, Timo Schick, and Hinrich Schtze. 2023. Semantic-oriented unlabeled priming for large-scale language models. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 32–38.

Anagh Malik. 2021. Strategies for the iterated prisoner's dilemma. *arXiv preprint arXiv:2111.11561*.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. 2023. Alympics: Language agents meet game theory. *arXiv preprint arXiv:2311.03220*.

Jennifer Dodorico McDonald. 2008. Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire*, 1(1):1–19.

Richard L Michalski and Todd K Shackelford. 2010. Evolutionary personality psychology: Reconciling human nature and individual differences. *Personality and Individual Differences*, 48(5):509–516.

Eladio Montero-Porras, Jelena Grujić, Elias Fernández Domingos, and Tom Lenaerts. 2022. Inferring strategies from observations in long iterated prisoner's dilemma experiments. *Scientific reports*, 12(1):7589.

John Nash. 1951. Non-cooperative games. *Annals of Mathematics*, 54(2).

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.

Matjaž Perc. 2016. Phase transitions in models of human cooperation. *Physics Letters A*, 380(36):2803–2808.

Emmanuel M Pothos, Gavin Perry, Philip J Corr, Mervin R Matthew, and Jerome R Busemeyer. 2011. Understanding cooperation in the prisoner's dilemma game. *Personality and Individual Differences*, 51(3):210–215.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Brent W Roberts and Avshalom Caspi. 2001. Personality development and the person-situation debate: It's déjà vu all over again. *Psychological Inquiry*, 12(2):104–109.

Brent W Roberts and Daniel Mroczek. 2008. Personality trait change in adulthood. *Current directions in psychological science*, 17(1):31–35.

Jose Gregorio Ferreira De Sá, Andreas Kaltenbrunner, Jacopo Amidei, and Rubén Nieto. 2024. How well do simulated populations with GPT-4 align with real ones in clinical trials? the case of the EPQR-a personality test. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

Mustafa Safdari, Greg Serapio-García, 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*.

Rajalakshmi Sivanaiah, Abrit Pal Singh, Aviansh Gupta, and Ayush Nanda. 2024. Agent enhancement using deep reinforcement learning algorithms for multiplayer game (slither. io). *International Journal of Next-Generation Computing*, 15(3).

Reiji Suzuki and Takaya Arita. 2024. An evolutionary model of personality traits related to cooperative behavior using a large language model. *Scientific Reports*, 14(1):5989.

Isabel Thielmann, Giuliana Spadaro, and Daniel Balliet. 2020. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological bulletin*, 146(1):30.

Eric Van Dijk and Carsten KW De Dreu. 2021. Experimental games and social decision making. *Annual Review of Psychology*, 72(1):415–438.

Gabriella Vigliocco, Lotte Meteyard, Mark Andrews, and Stavroula Kousta. 2009. Toward a theory of semantic representation. *Language and Cognition*, 1(2):219–247.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*.

Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. 2025. Will systems of llm agents cooperate: An investigation into a social dilemma. *arXiv preprint arXiv:2501.16173*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Qin Yang, Yi Tang, and Dehua Gao. 2024. Agent-based evolutionary game dynamics uncover the dual role of resource heterogeneity in the evolution of cooperation. *Journal of Theoretical Biology*, 595:111952.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. 2024. Human vs. generative ai in content creation competition: Symbiosis or conflict? *arXiv preprint arXiv:2402.15467*.

Jianlei Zhang, Xiaojie Chen, Chunyan Zhang, Long Wang, and Tianguang Chu. 2010. Elimination mechanism promotes cooperation in coevolutionary prisoner's dilemma games. *Physica A: Statistical Mechanics and its Applications*, 389(19):4081–4086.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024. Competeai: Understanding the competition dynamics of large language model-based agents. In *Forty-first International Conference on Machine Learning*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Prompts

A.1 The Prisoner's Dilemma

To maximize control variables, we keep the prompts in the Prisoner's Dilemma (except the mutation process) consistent with Suzuki and Arita (2024) as much as possible.

Game Rules Answer the questions, please act as if your personality trait is ${personality_description}. This is an experimental study of strategy choices in different situations. You and your partner have an opportunity to earn money based on individual decisions. You and your partner will independently choose one of two strategies: 'Strategy A' or 'Strategy B'. Your earnings will depend on the combination of your choice and your partner's choice. If both of you choose 'Strategy A', each of you earns $4, which is a mutually intermediate amount. If you choose 'Strategy A', and your partner chooses 'Strategy B', you earn $0, which is the smallest, and your partner earns $5, which is the largest. If you choose 'Strategy B', and your partner chooses 'Strategy A', you earn $5, which is the largest, and your partner earns $0, which is the smallest. If both of you choose 'Strategy B', each of you earns $1, which is not zero but smaller than $4. For example: If you choose 'Strategy A' and your partner also chooses 'Strategy A', both of you will earn $4 each. However, if you choose 'Strategy A' and your partner chooses 'Strategy B', you will earn nothing, while your partner will earn $5. Your goal is to maximize the total money you receive through repeated rounds of decisions regardless of the other's money. Not a combined earning but your own money. Predicting your partner's choice and considering your personality trait, which strategy do you choose in the next round? Please answer using the format 'I choose Strategy *.' No further mentions are allowed, but you may ask the reason for your choice later.

Initial Personality Description

1. Pursues personal gain consistently, neglecting mutual or group benefits entirely.
2. Open to team efforts, but self-interest frequently overrides collective goals.
3. Recognizes cooperation's value but often demands trust before committing fully.
4. Balances between individual needs and team benefits based on situations.
5. Values collaboration, though, retains a watchful eye for possible betrayals.
6. Favors group outcomes, believes in shared growth, occasionally sets limits.
7. Commits wholly to teamwork, placing the group's interests above personal ones.

Prompting agents to make a choice Predicting your partner's choice and considering your personality trait, which strategy do you choose in the next round? Please answer using the format 'I choose Strategy *.' No further mentions are allowed, but you may ask the reason for your choice later.

Showing results to agents You chose ${strategy_name_1} and your partner chose ${strategy_name_2} in the ${round_count} round. So, you got $${points_1} and your partner got $${points_2} in the ${round_count} round.

The mutation process of personality Answer the questions. The following text describes the character of a person. ${personality_description} Please rephrase the description of a personality trait within 10 words. Your answer starts with 'Rephrased text:'

A.2 BFI Evaluation

The `${self_description}` in our BFI evaluation prompts is based on the inventory proposed by Johnson (2014).

Prompting agents to give self-evaluation `You will be asked a question, please please act as if your personality trait is ${personality_description} when answering the question. Do you think the description "You ${self_description}" applies to you? Please rate your level of agreement on a scale from 0 to 6: 0 means completely disagree, 1 means strongly disagree, 2 means a little disagree, 3 means neither agree nor disagree, 4 means little disagree, 5 means strongly agree, 6 means completely disagree. Please ONLY output the number and DO NOT add any additional field or line break to your response!`

B Word Lists for Word Frequency Statistics

B.1 Words with Behavioral Semantic

- Collaboration (W-Col): altruistic, cohesion, collaboration, collective, communal, community, cooperative, generous, group, helpful, others, selfless, share, supportive, team, teamwork, team-centered, team-oriented, unite, unity
- Defection (W-Def): boundary, egocentric, independence, independent, individual, individualistic, limits, narcissistic, own, personal, self-absorbed, self-centered, self-focused, self-interest

B.2 Privative Words

Only the verb prototype is listed.

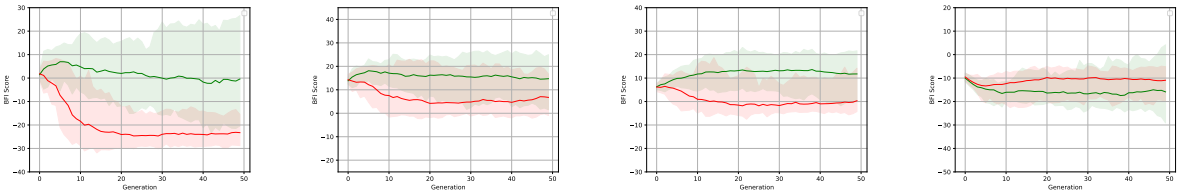
- above, disregard, ignore, neglect, over, rather than, undervalue

C Variation of Personality Metrics

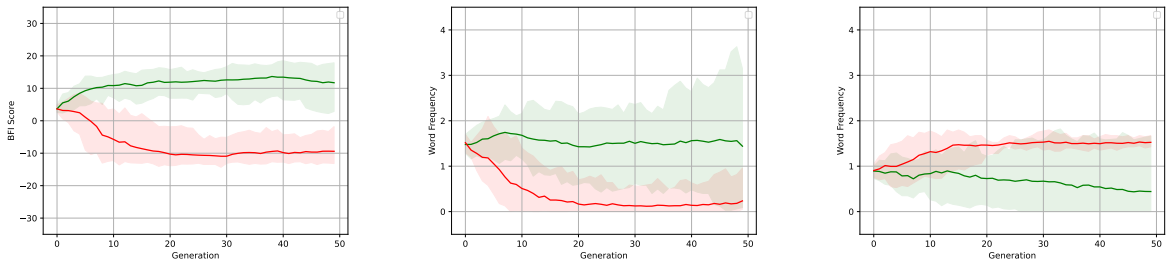
Figure 5 and Figure 6 show the variation of BFI scores and word frequency in Group-Col and Group-Def. The separation effects of $\mathrm{BFI}_{\mathrm{A}}(t)$ and $\mathrm{WF}_{\mathrm{Col}}(t)$ are significant under both GLM-4-Air and Deepseek-V3, while the separation of $\mathrm{BFI}_{\mathrm{O}}(t)$ is only significant under GLM-4-Air.

D AI Assistance Disclosure

The authors acknowledge the use of AI-assisted tools for non-content aspects of this paper. As non-native English speakers, we employed AI tools to refine the linguistic expression of certain sections, particularly focusing on enhancing readability. However, we emphasize that all research ideas, methodological designs, and coding implementations were developed without AI involvement.

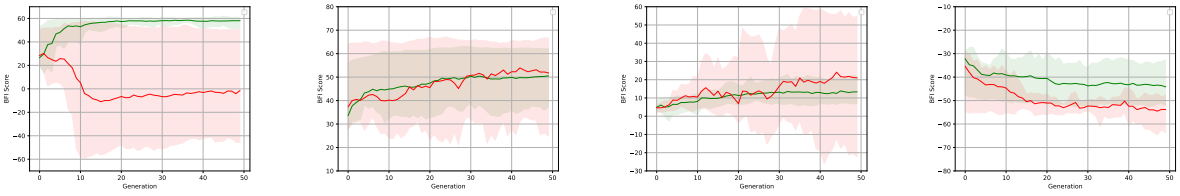


(a) $\mathrm{BFI_A}(t)$ (b) $\mathrm{BFI_C}(t)$ (c) $\mathrm{BFI_E}(t)$ (d) $\mathrm{BFI_N}(t)$

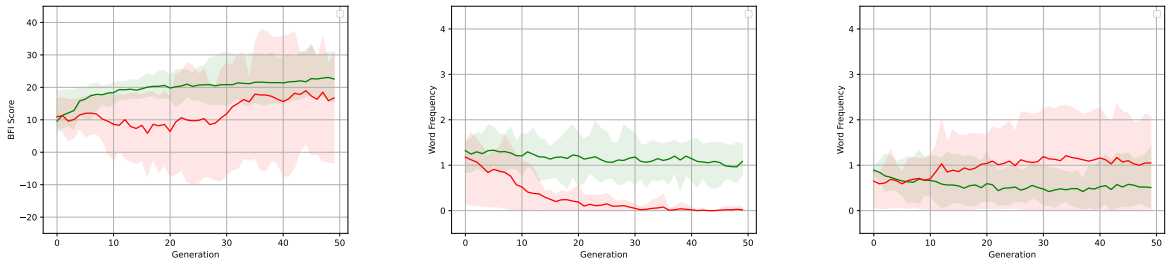


(e) $\mathrm{BFI_O}(t)$ (f) $\mathrm{WF_{Col}}(t)$ (g) $\mathrm{WF_{Def}}(t)$

Figure 5: General variations of personality metrics in Group-Col (green) and Group-Def (red) among 50 experiments. Agents use GLM-4-Air as base LLM ($p_m = 0.05, \tau = 0.95$). The solid line represents the average score, and the shaded area shows the 95% distribution range of experimental results.



(a) $\mathrm{BFI_A}(t)$ (b) $\mathrm{BFI_C}(t)$ (c) $\mathrm{BFI_E}(t)$ (d) $\mathrm{BFI_N}(t)$



(e) $\mathrm{BFI_O}(t)$ (f) $\mathrm{WF_{Col}}(t)$ (g) $\mathrm{WF_{Def}}(t)$

Figure 6: General variations of personality metrics in Group-Col (green) and Group-Def (red) among 16 experiments. Agents use Deepseek-V3 as base LLM ($p_m = 0.05, \tau = 0.95$). The solid line represents the average score, and the shaded area shows the 95% distribution range of experimental results.