Predictive Validity of LLM Personality Reports: a Case Study on Classic Economic Games

Yin Jou Huang Rafik Hadfi
Graduate School of Informatics, Kyoto University, Japan
huang@nlp.ist.i.kyoto-u.ac.jp, rafik.hadfi@i.kyoto-u.ac.jp

Abstract

Personality assessment methods used in psychology are now being applied to study LLM agents. However, the behavioral validity of such methods remains unclear. One commonly used approach relies on personality reports. In this work, we evaluate the predictive validity of LLM personality reports by testing whether self-report and observer-report profiles can predict strategic behavior in two-player games. Our results show that observer reports are more predictive in coordination-heavy settings, while self-reports are more predictive in high-conflict games. We also find that combining both paradigms yields the strongest overall predictive power, suggesting that they capture complementary personality signals. We adopt predictive validity as a principled framework for evaluating LLM personality assessment methods.

1 Introduction

Large language model (LLM) agents are increasingly deployed in interactive social settings, including mental health support, or education [1, 2]. As these agents see wider deployment, there has been a growing interest in analyzing their behavioral tendencies through personality assessments. The goal is to improve human–AI interaction, alignment, and predictability.

Most existing approaches of LLM personality assessment adapt human personality inventories such as the Big Five personality theory [3]. In psychology, we generally encounter two contrasting paradigms of self-report and observer-report. According to the self-report paradigm, an LLM agent directly rates standardized personality items to produce an introspective profile. This approach is simple to deploy but suffers from reliability issues, prompt sensitivity, and systematic biases [4, 5, 6]. In contrast, the observer-report paradigm infers the personality of an LLM agent from externally observable behaviors, elicited through interaction with other agents [6]. Existing work has also shown that observer-reports diverge substantially from self-reports, suggesting that the two paradigms capture different personality traits of LLM agents.

A fundamental challenge for personality assessment is that personality is a latent construct with no direct ground truth. In psychology, the validity of a personality assessment method is established through **predictive validity**, which is the extent to which personality measures can predict consequential behaviors and real-world outcomes such as job performance, relationship quality, and health-related decisions [7]. Predictive validity is widely regarded as the gold standard for evaluating assessments of latent traits like personality. However, it remains unclear whether the personality profiles of LLM agents obtained via self-report or observer-report can meaningfully predict behavioral patterns, and not just capture surface-level textual regularities in generated texts.

To address this gap, we examine the predictive validity of LLM personality assessments by evaluating if self-reports and observer-reports can predict agent behavior in strategic decision-making scenarios. We specifically employ a set of classic two-player economic games: Pure Coordination, Stag Hunt, Battle of the Sexes, Hawk–Dove, and the Prisoner's Dilemma [8, 9]. These games have long served as behavioral probes in economics, game theory and psychology as they provide a solid and controlled testbed for studying behavioral tendencies like cooperation, risk-taking, and reciprocity.

Using such games, we systematically compare the predictive power of self-report and observer-report personality profiles at two levels of analysis. At the macro level, we examine how well personality measures predict ag-

gregate game outcomes. At the micro level, we analyze whether personality traits predict fine-grained action patterns such as cooperation, risk-taking, forgiveness, reciprocity, and behavioral switching. Our results show that observer-reports and self-reports capture distinct and complementary personality signals. Observer-reports are more predictive of socially manifested actions such as cooperation, forgiveness, and risk-taking, while self-reports better predict internally driven strategy-adjustment behaviors such as reciprocity and switching.

2 LLM Personality Assessment

We now introduce self-report and observer-report paradigms for assessing LLM personality. Both methods use the same standardized personality questionnaires.

2.1 Big Five Personality Framework

We adopt the Big Five personality framework [3], consisting of Openness (OPE), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU). The Big Five theory provides the basis for configuring LLM agents and subsequent personality assessments.

LLM agent with synthetic personality We configure the LLM agents with synthetic Big Five personality profiles [10, 11]. Each agent is instantiated with a personality configuration specified through personality-describing adjectives corresponding to the five dimensions, injected into the system prompt to condition the agent's behavior.

Personality questionnaire We measure the personality profile of LLM agents using a 50-item questionnaire derived from the International Personality Item Pool (IPIP) [3]. Each item is a declarative statement (e.g., "I sympathize with others' feelings") and is mapped to exactly one of the five trait dimensions. Following prior work on LLM personality assessment, each questionnaire item is presented independently, and responses are collected on a 5-point Likert scale [10, 11]. Trait scores are computed by aggregating item-level responses within each dimension, yielding a 5-dimensional personality vector.[1)]

2.2 Self-report

In the self-report paradigm, an LLM agent is prompted to evaluate its own personality by responding to standardized questionnaire items such as "I sympathize with others' feelings" on a Likert scale based on how accurately it believes the statement reflects itself. The responses are aggregated to give a personality profile across predefined trait dimensions. Conceptually, self-report reflects the agent's internal self-description, relying on its ability to interpret personality concepts, map them onto its own behavioral tendencies, and produce a coherent introspective judgment.

While this approach is appealing due to its simplicity and efficiency, concerns remain about its reliability and the potential biases [4, 6]. Since self-reports rely on metacognitive reasoning, the resulting personality profiles may be influenced by factors such as instruction framing and exposure to personality metadata, causing it to deviate from the agent's actual behavioral patterns [6]. Consequently, self-reported personality should be interpreted as a declaration of how the agent perceives itself, rather than as a direct measurement of its behavioral tendencies.

2.3 Observer-report

The observer-report paradigm is proposed to address the limitations of self-report methods by inferring personality profile from externally observable behavior [6]. A target LLM agent interacts with multiple observer agents across a range of dialogue scenarios designed to elicit personality-relevant behaviors. Each observer agent evaluates the target agent using the same standardized personality questionnaire. Instead of rating how well each statement describes itself, the observer rates how well the statement describes the target agent's behavior. Observer-report captures the agent's perceived personality manifested in its behavior, language style, and interaction dynamics. Notably, there is a systematic divergence between observer-reports and self-reports of LLM agents, suggesting that the two paradigms capture different aspects of personality-related signals.

3 Economic Games

We adopt a game-theoretic approach in which each subject LLM agent engages in a 2-player economic game with another opponent agent [8].

3.1 Game Types

We evaluate the decisions using five classic two-player economic games: Battle of the Sexes (BS), Pure Coordination (PC), Hawk-Dove (HD), Prisoner's Dilemma (PD),

1) The list of 50 items and the scoring schemes can be found at https://ipip.ori.org/newBigFive5broadKey.htm.

and Stag Hunt (SH) [8]. Each game is defined using a payoff matrix that encodes the preferences of the agents. Refer to the appendix for the details.

The five games can be arranged along a spectrum of increasing conflict. PC involves no conflict, as players have identical preferences and simply need to match actions. SH introduces minimal conflict through risk asymmetry, where mutual cooperation yields the highest payoff (both choosing the first action 'stag'), but a safe alternative (the second action 'hare') creates tension around trust [12]. BS adds distributional conflict while retaining the coordination motives. In this case, players want to match actions but prefer different actions. HD shifts toward anti-coordination, where players benefit from mismatched actions but face costly mutual aggression. Finally, PD represents maximum conflict with no coordination motive (choosing the first action), where each player has a dominant strategy to defect (the second action) [8].

Simulation based on LLM Agents Herein, payoff matrices are translated into textual instructions and presented to LLM agents [13]. We randomly pair the LLM agents configured in Section 2.1. Each pair of agents plays a repeated game of 10 rounds, for all five types of games. In each round, an agent observes the opponent's previous action before each decision. We record an agent's cumulative reward as the game outcome, yielding a 5-dimensional outcome vector $U \in [0, 9]^5$ across the five games.

3.2 Action Types

To analyze fine-grained behavioral tendencies beyond aggregate payoffs, we categorize each agent's round-level decisions into five interpretable action types. To **cooperate** is to promote mutually beneficial outcomes or successful coordination between players. Specifically, the first action (cooperate) of the Prisoner's Dilemma (PD) game, the first action (stag) of the Stag Hunt (SH), and choosing the same action as the opponent's action in the Pure Coordination (PC) game are defined as cooperation actions. The other action is thus defined as the defection action. To **forgive** is to maintain cooperation despite the opponent had defected in the previous round. This applies to the PD, SH, and PC games. Forgiveness captures tolerance to short-term exploitation or miscoordination. An agent **reciprocates** if their current action is identical to the opponent's action in the previous round. This definition applies to all game types, and captures reactive strategies such as tit-for-tat [14]. To **risk** is to be exposed to higher potential loss under uncertainty about the opponent's behavior. We define the risk of an action using the maximum regret criterion [15]. Maximum regret is the largest difference between the payoff that could have been achieved by choosing the best response in hindsight and the payoff obtained by the chosen action, across all possible opponent actions. If one of the agent's two possible actions has higher maximum regret, that action is defined as a risk-taking action. For example, choosing the first action 'stag' in SH game, second action 'hawk' in HD game, first action 'cooperate' in PD game, and the second action (opponent's preferred outcome) in BS game are risky actions. Refer to the appendix for the formal definition of risk. Finally, to **switch** is to act differently from the agent's own action in the previous round. This definition applies to all game types and captures exploratory or stationary behavior.

For each LLM agent, we record the frequency of each of the above action types, yielding a 5-dimensional action vector $A \in [0, 1]^5$ across the five action types.

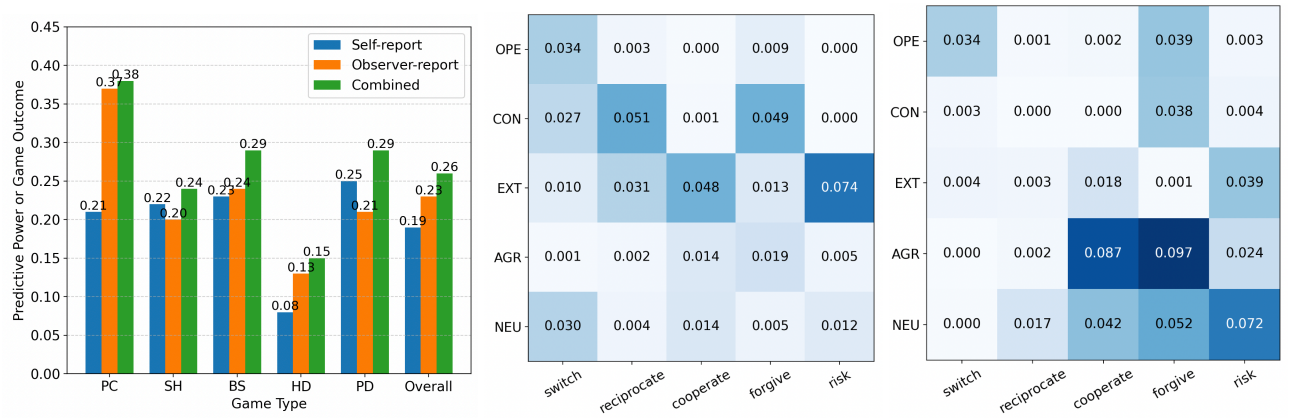
4 Predictive Behavioral Analysis

4.1 Predictive Power of Game Outcome

To assess the predictive validity, we train linear regression models that map personality vectors to game outcomes U. We compare three predictors: self-report, observer-report, and a combined predictor that concatenates both personality vectors. Predictive performance is measured using the coefficient of determination (R^2), reported in Figure 1a.

Overall, observer-reports slightly outperform self-reports across games, while the combined predictor consistently achieves the highest predictive power. This indicates that observer and self-reports capture unique information that the other cannot provide, and complement each other.

Predictive power varies substantially across game types. Games with strong coordination incentives, such as Pure Coordination (PC), can be predicted more from observer-reports. In these settings, behavior is highly responsive to social signaling, and externally observable interaction patterns provide informative cues. In contrast, high-conflict games such as the Prisoner's Dilemma (PD) show weaker gains from observer-reports and relatively stronger contri-



(a) Predictive Power of economic game outcome. (b) Predictive power of self-report. (c) Predictive power of observer-report.

Figure 1: Predictive power of self-report and observer-report personality across outcomes and actions.

butions from self-reports. Since defection is the dominant strategy, communication and perceived intent play a limited role, reducing the value of externally inferred personality signals. Games with anti-coordination or mixed incentives, particularly Hawk–Dove (HD), exhibit the lowest predictive power across all predictors. In such environments, strategic unpredictability and context-dependent behavior limit the explanatory role of stable personality traits. Battle of the Sexes (BS) and Stag Hunt (SH) fall between these extremes, showing modest predictability consistent with their multiple equilibria and reliance on situational factors.

4.2 Predictive Power of Agent Action

To complement the macro-level analysis of game outcomes, we now examine whether personality dimensions predict fine-grained action patterns at the micro level. Specifically, we analyze how self-report and observer-report personality profiles predict the frequency with which agents cooperate, reciprocate, forgive, risk, and switch. Figure 1b and 1c report the predictive power of each Big Five dimension for these five action types.

We observe a clear asymmetry between the predictive power pattern of self-report and observer-report. Self-reports tend to be more informative for predicting internally driven or strategy-adjustment behaviors, such as switch and reciprocate. In contrast, observer-reports provide stronger and more consistent predictive signals for socially oriented actions, including cooperate, forgive, and risk. This mirrors the macro-level finding that observer-reports excel in interaction- and coordination-heavy settings.

The difference between self-report and observer-report is also shown in the distribution of their predictive dimension. The predictive power of self-report is concentrated in EXT and CON dimensions. EXT emerges as the most predictive dimension, particularly for risk-taking ($R^2 = 0.074$) and cooperation ($R^2 = 0.048$), aligning with similar tendencies in human. CON predicts reciprocity ($R^2 = 0.051$) and forgiveness ($R^2 = 0.049$). This links to self-perceived focus on rule-following behavior. For observer-reported personality, we see a different pattern that concentrates on AGR and NEU dimensions. AGR is the dominant predictor, strongly linked to cooperation ($R^2 = 0.087$) and forgiveness ($R^2 = 0.097$). Accordingly, pro-social tendencies manifested during interaction indicate cooperative and forgiving actions. On the other hand, NEU has a high predictive power of forgiveness ($R^2 = 0.072$) and cooperation ($R^2 = 0.042$). This implies that emotionally unstable behaviors that are visible to observers signal how agents respond to conflict.

Overall, these findings suggest that self-report and observer-report capture complementary personality signals at the action level, supporting the combined predictive gains in Section 4.1.

5 Conclusion

In this work, we demonstrate that LLM personality assessments exhibit predictive validity for strategic behaviors in economic games. Both macro- and micro-level analysis of predictive power show that self-report and observer-report capture complementary behavioral signals. The findings support predictive validity as a principled criterion for evaluating latent traits of LLMs.

Acknowledgment

This work was supported by JST ACT-X Grant Number JPMJAX23CP and JSPS Kakenhi Grant Number JP23K28145.

References

[1] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. **arXiv preprint arXiv:2307.11991**, 2023.

[2] Y. Hicke, A. Agarwal, Q. Ma, and P. Denny. Chata: Towards an intelligent question-answer teaching assistant using open-source llms. **arXiv preprint arXiv:2311.02775**, 2023.

[3] L. R. Goldberg. The development of markers for the big-five factor structure. **Psychological Assessment**, Vol. 4, No. 1, pp. 26–42, 1992.

[4] A. Gupta, X. Song, and G. Anumanchipalli. Self-assessment tests are unreliable measures of llm personality. In **Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP**, 2023.

[5] Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. Do personality tests generalize to large language models? In **Socially Responsible Language Modelling Research**, 2023.

[6] Yin Jou Huang and Rafik Hadfi. Beyond self-reports: Multi-observer agents for personality assessment in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 21086–21101, Suzhou, China, November 2025. Association for Computational Linguistics.

[7] Daniel J Ozer and Veronica Benet-Martinez. Personality and the prediction of consequential outcomes. **Annu. Rev. Psychol.**, Vol. 57, No. 1, pp. 401–421, 2006.

[8] Robert Gibbons. **Game theory for applied economists**. Princeton University Press, 1992.

[9] Rafik Hadfi and Yin Jou Huang. Personality-aware multiagent large language models for strategic interactions in polymatrix games. In **Proceedings of the 13th International Conference on Human-Agent Interaction**, HAI '25, p. 521–523, New York, NY, USA, 2026. Association for Computing Machinery.

[10] G. Serapio-García, et al. Personality traits in large language models. **arXiv preprint arXiv:2307.00184**, 2023.

[11] Yin Jou Huang and Rafik Hadfi. How personality traits influence negotiation outcomes? a simulation based on large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 10336–10351, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] John Hartley, Conor Brian Hamill, Dale Seddon, Devesh Batra, Ramin Okhrati, and Raad Khraishi. How personality traits shape llm risk-taking behaviour. In **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 21068–21092, 2025.

[13] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. **Nature Human Behaviour**, pp. 1–11, 2025.

[14] Robert Axelrod and William D Hamilton. The evolution of cooperation. **science**, Vol. 211, No. 4489, pp. 1390–1396, 1981.

[15] Yoav Shoham and Kevin Leyton-Brown. **Multiagent systems: Algorithmic, game-theoretic, and logical foundations**. Cambridge University Press, 2008.

A The Economic Games

We look into five classic two-player games: Pure Coordination (PC), Stag Hunt (SH), Prisoner's Dilemma (PD), Battle of the Sexes (BS), and Hawk-Dove (HD) [8]. The corresponding payoff matrices are described in equations (1).

$$
\begin{aligned}
M_{\text{PC}} &= \begin{pmatrix} 1,1 & 0,0 \\ 0,0 & 1,1 \end{pmatrix} & M_{\text{SH}} &= \begin{pmatrix} 9,9 & 0,8 \\ 8,0 & 7,7 \end{pmatrix} \\
M_{\text{PD}} &= \begin{pmatrix} 3,3 & 0,4 \\ 4,0 & 1,1 \end{pmatrix} & M_{\text{BS}} &= \begin{pmatrix} 2,1 & 0,0 \\ 0,0 & 1,2 \end{pmatrix} \\
& & M_{\text{HD}} &= \begin{pmatrix} 2,2 & 1,3 \\ 3,1 & 0,0 \end{pmatrix}
\end{aligned}
\tag{1}
$$

B Risk Quantification

The risk mentioned in Section 3.2 is formally defined as the maximum regret criterion [15]. An agent i's risk, or maximum regret, for playing an action a_i is defined as in equation (2).

$$
risk(a_i) = \max_{a_{-i} \in A_{-i}} \left([\max_{a_i' \in A_i} u_i(a_i', a_{-i})] - u_i(a_i, a_{-i}) \right) \tag{2}
$$

Here, u_i is the payoff function of i as defined in the payoff matrices, and A_i is its action set containing two actions. Equation (2) represents the amount that i loses by playing a_i rather than playing its best response to opponent's action a_{-i}, assumed to make this loss as large as possible. An agent is considered risk-taking when it selects actions with high maximum regret, accepting greater potential losses in pursuit of higher payoffs.