

Words or Deeds? How LLMs Predict Users’ Decisions in Human-Agent Dialogue

Baiqiao Zhang[♠], Shuai Ma[◇], Zhifeng Liao[♠], Chao Zhou[◇], Jianshu Wang[◇], Yuhao Zhao[♠],
Chaoyue Zhang[♠], Tongjun Si[♠], Yuying Tang[♠], Xiangxian Li[♠], Yong-Jin Liu[♡], Yulong Bian^{♠*}

[♠] The Hong Kong University of Science and Technology, Hong Kong SAR, China

[◇] Institute of Software, Chinese Academy of Sciences, Beijing, China

[♠] Shandong University, Weihai, China

[♡] Tsinghua University, Beijing, China

bzhangcw@cse.ust.hk, {mashuai, zhouchao}@iscas.ac.cn, yuying.tang@connect.ust.hk, jeffwang825@henu.edu.cn,

{202200810061, 202400800660, 202400800527, 202400800539}@mail.sdu.edu.cn,

{xiangxianli, bianyulong}@sdu.edu.cn, liuyongjin@tsinghua.edu.cn

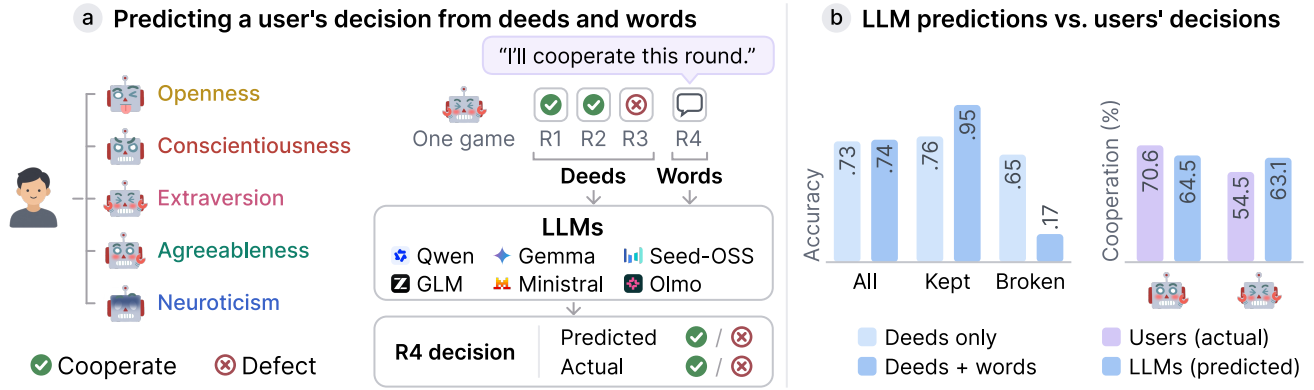


Figure 1: LLM predictions varied with users’ stated decisions but barely varied with the agent users faced, although users’ behavior did. (a) Each user played a six-round game with each of five agents, each prompted to express a dominant Big Five trait, and could talk with the agent before deciding to cooperate or defect in every round. Eight open-weight LLMs predicted each decision from earlier decisions (deeds) and the round’s dialogue (words). (b) Left: adding dialogue barely changed accuracy on rounds with a stated decision, but raised it on kept statements and lowered it on broken ones. Right: users cooperated more with the Agreeableness than with the Extraversion agent, whereas predictions from all decisions and dialogue barely differed.

Abstract

Web systems increasingly predict users’ decisions from what users say, yet when interests partly conflict, users may promise to cooperate to win the other side’s cooperation and then not do so. Systems that rely on such promises, such as sales or service agents deciding what to offer next, may grant concessions to users who then do not keep their promises, at a cost to the party each agent represents. We test how LLMs combine users’ statements with decision records to predict users’ decisions, using 1,260 rounds in which 42 users played a repeated prisoner’s dilemma-style game with five agents, each prompted to express a dominant Big Five trait, and could talk with the agent before every decision. Across eight open-weight LLMs, predictions from decision records separated cooperation from defection well (AUC = .80), and adding dialogue

brought no detectable gain. With dialogue, predictions agreed with users’ statements more often, and errors concentrated on rounds in which users broke their statements: accuracy on rounds with a statement barely changed (.73 to .74), while it rose on kept statements (.76 to .95) and fell sharply on broken ones (.65 to .17). Predictions weighted statements more heavily than users’ behavior warranted: when users stated a decision they had not yet made in that game, the predicted probability of keeping it rose from .13 to .69, whereas users kept .40 of such statements. Users’ decisions also varied with the agent they faced, but predictions hardly reflected this: users cooperated more with the Agreeableness agent than with the Extraversion agent (70.6% vs. 54.5%), whereas predicted cooperation was nearly identical (64.5% vs. 63.1%). Users also cooperated more with an agent when they themselves scored higher on that agent’s trait (e.g., agreeable users with the Agreeableness agent), whereas predicted cooperation barely changed with these scores. Dialogue-based predictions should therefore weigh users’ statements by their earlier decisions and the agent they face.

*Corresponding author: Yulong Bian.



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

© 2026 Copyright held by the owner/author(s).

CCS Concepts

• **Information systems** → **Personalization**; • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Natural language processing*.

Keywords

User Modeling, Human-AI Interaction, Decision Making

1 Introduction

Web systems predict what users will do next. Recommendation and ranking have long drawn these predictions from past behavior, such as clicks and purchases [7, 9, 20]. Conversational interaction adds a more direct source, because users often say what they will do before they act. As LLM-based agents enter Web platforms, users talk with sales and service agents that represent another party and then decide whether to accept an offer or keep an agreement [21, 30]. When a user tells such an agent “I will accept this offer,” the dialogue appears to contain the answer to the prediction task.

User modeling has built the means to use such language. Conversational recommenders model users from the current dialogue [8, 11, 36]; LLMs predict individuals’ choices from interviews [18], from earlier choices in psychology experiments [1], from play in persuasion games [22] and after conversations [19], and stand in for users when conversational agents are evaluated [15, 33–35]. Of the language available to a predictor, a user’s own statement about the predicted action is the most direct evidence. A statement fixes the action the user names, but not how likely the user is to take it.

This gap matters most in mixed-motive interactions, where both sides gain from cooperation while their interests partly conflict: there a statement expresses a plan and also seeks to influence the other side. Promises raise cooperation in experimental games, yet about a quarter of them are broken [2, 27], and people broke more than half of their cooperation agreements with LLM agents [29]. A predictor therefore has to decide how far a statement should move its prediction, above all when the statement is inconsistent with the user’s earlier decisions: ignoring it can miss real changes of plan, and relying on it heavily can overestimate how often it is kept. Overall accuracy does not reveal which error a predictor makes: a predictor that relies on statements gains accuracy on kept statements and loses it on broken ones, and the two can cancel. Whether dialogue improves prediction must therefore be judged against what users do later, separately for kept and broken statements.

A statement is also made to a particular counterpart, and the action is taken toward it. A counterpart’s conduct and persona change people’s decisions [3, 14, 29], and how people respond to an agent’s personality depends on their own: people were more attracted to a computer whose personality matched theirs [16], and human-AI teamwork depended on how the human’s and the AI’s traits were paired [13]. Users’ decisions may therefore change with the agent’s persona and with how that persona matches the user’s own traits. A predictor that is accurate over all agents together can still be biased with particular agents or user-agent pairs.

Existing studies cannot settle either issue, as their data were collected for other questions. In choice-prediction studies, users receive messages from a bot [22], talk without stating the predicted

action [19], or play one-shot games with no record of earlier behavior [6]. User-simulator evaluations score generated conversation turns, so no later action can contradict what a simulated user says, and they compare responses to different assistants after pooling over users [15]. No study links a user’s statement to that user’s earlier decisions, the counterpart, and the later action.

We therefore ask three research questions.

- **RQ1:** How do LLM predictors combine users’ stated decisions, what users say they will do in the current round, with the record of earlier decisions, and is their reliance on statements supported by how often statements are kept?
- **RQ2:** Do users’ decisions and word-keeping vary with the agent persona, and do the predictions vary in the same way?
- **RQ3:** Do users’ decisions vary with their own personality, measured as *trait compatibility*, the user’s score on the agent’s defining trait, and do the predictions vary in the same way?

Answering them requires statements and actions recorded separately, repeated decisions that form a record against which to compare a statement, and counterparts that change while actions and payoffs stay the same. A repeated prisoner’s-dilemma-style game with communication before each decision meets these requirements. We use the Personality-1260 dataset [31], in which 42 users played six rounds with each of five GPT-4o agents, each prompted to express a dominant Big Five trait, talking with the agent before every decision (Figure 1a). Eight open-weight LLMs predicted each decision from 19 combinations of the decision record (both players’ earlier decisions, the *deeds*) and the dialogue (what users said, the *words*). We evaluate predictions separately for kept and broken statements and for those inconsistent with users’ earlier decisions, measure the effect of one inserted statement, and compare them with users’ decisions across agents and trait compatibility.

Predictions from the decision record separated cooperation from defection well ($AUC = .80$), and adding dialogue brought no detectable gain. Adding the round’s dialogue left accuracy on rounds with a stated decision almost unchanged (.73 to .74), because it raised accuracy where users kept their statement (.76 to .95) and lowered it where they broke it (.65 to .17; Figure 1b, left). When users stated a decision they had not yet chosen in that game, they kept it at a rate of .40, while the predicted probability of keeping it rose from .13 without the dialogue to .69 with it: the predictions moved beyond what the statement warranted. Users cooperated in 70.6% of rounds with the Agreeableness agent and 54.5% with the Extraversion agent and kept more of their statements with the former; predicted cooperation was 64.5% and 63.1% (Figure 1b, right). Users also cooperated more as trait compatibility rose, whereas the predictions barely changed and fell increasingly below users’ cooperation. Pooled over agents, the prediction bias was near zero.

This paper makes four contributions:

- (1) An evaluation design that ties a predictor’s use of dialogue to separately recorded actions, split by whether statements were kept, whether they were inconsistent with users’ earlier decisions, and the agent the user faced.
- (2) Evidence that eight LLM predictors relied on stated decisions beyond how often they were kept.
- (3) Evidence that the predictions missed how users’ behavior varied with the agent persona and trait compatibility, and that pooled metrics concealed both errors.

- (4) Design implications for dialogue-based user modeling: how much weight a user’s statement receives should depend on the user’s earlier decisions and on the agent, and predictions should be checked against what users later do with each agent.

2 Related Work

2.1 Prediction from Dialogue and History

Studies of decision prediction have used behavioral history, conversational information [24], or both. Nay and Vorobeychik [17] model cooperation in repeated games using the game structure and both players’ previous actions. In language-based persuasion games, Shapira et al. [22] find that LLM-generated data support better human-choice prediction when they capture history-dependent behavior beyond the sentiment of the messages users receive. Dialogue can also provide context without revealing the user’s intended choice: Pilli and Nallur [19] predict decisions after cognitively demanding conversations, but replacing users’ replies with random text changes accuracy little. In these settings, dialogue provides context for a choice without containing the user’s statement of that choice. Gonzalez-Fernandez et al. [6] examine such statements more directly by predicting strategic choices from pre-play communication. In their three games, GPT-5 achieved higher mean accuracy than human forecasters, but was much less accurate for defection than cooperation. Their one-shot settings, however, provide no repeated decision history against which a current statement can be assessed. Repeated interactions allow a further question: does a predictor interpret a current statement in light of the speaker’s earlier choices and record of keeping statements?

2.2 Stated Decisions as Behavioral Evidence

A stated decision can be informative without guaranteeing that the stated action will occur. Promises can encourage trustworthy behavior [2], and their wording matters: conditional or implicit statements in a high-stakes game show were associated with less cooperation than explicit, unconditional ones [26]. Yet even expressed agreement can precede a different action. In online donation dialogues, some participants who agreed to donate subsequently gave less or nothing [28]. We use *stated decisions* to include expressions of intended action beyond explicit promises, without assuming that they will be kept. Recognizing that statements are informative does not ensure that an observer uses them appropriately. Chen and Houser [4] found that human observers distinguished promises from empty talk, but overestimated promise keeping and could not reliably distinguish kept from broken promises. This separates two questions: whether statements predict behavior, and whether a forecaster’s confidence in them matches observed behavior.

2.3 Counterparts and User-Agent Pairing

The behavior against which these forecasts are assessed also depends on the counterpart. de Melo et al. [5] found that emotional displays altered cooperation with agents using the same decision strategy. In conversational negotiation, agents’ willingness to compromise affects whether people reach agreements [3]; with LLM counterparts, both cooperation and agreement keeping vary across personas [29]. These findings make the agent’s conduct part of the context for predicting a user’s next action.

Responses to that conduct can further depend on the user. Personality similarity increased attraction to computers in Moon and Nass [16], whereas Isbister and Nass [10] found a preference for complementary interactive characters. More recently, Ju and Aral [13] found that human and AI personality pairings affected teamwork and performance. We examine how the user’s score on the agent’s defining trait relates to decisions and prediction bias, without assuming that higher compatibility implies greater cooperation.

For prediction, the question is whether forecasts reproduce such contextual differences. ConvApparel tests whether simulated conversational responses reproduce real users’ reactions to assistants of differing quality, finding that data-driven simulators adapt more faithfully than a prompted one [15]. We bring this evaluation question to separately recorded actions after dialogue, comparing forecasts with behavior across agent personas and user-agent pairings.

3 Experiment Design

We study offline prediction on recorded games: for each round, a predictor receives material recorded before the user’s decision and outputs the probability that the user cooperates. The agent played with the user; the predictor is a separate model, run afterwards, that only forecasts the user’s recorded decision.

3.1 Users, Agents and Game

We use the Personality-1260 dataset [31]. Forty-two participants from a Chinese university (21 female; mean age 22.1) completed the Chinese BFI-44 [12] and played one game with each of five agents in random order. In each of six rounds per game, user and agent may talk freely; then both decide whether to cooperate or defect, the agent’s decision being revealed only after the user’s. Mutual cooperation pays 2 points each, defecting against a cooperator 3 (the cooperator 0), and mutual defection 0 each. Talking was optional, and the dialogue was empty in 276 of the 1,260 rounds.

Each agent is GPT-4o prompted to be extremely high in one Big Five trait, to maximize its score, and to show the trait in dialogue and decisions without naming it. GPT-4o expressed personas set to the high end of a trait more faithfully than other models [23], and blind experts rated each agent 4.53–4.80 of 5 on its trait [31]. We call the agents by their *agent persona*: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism (N). Users cooperated in 64.5% of rounds and agents in 77.7% (Appendix F).

3.2 Stated Decisions

A *stated decision* is what the user said in the round’s dialogue about their own decision in that round. Two researchers independently labeled each dialogue, shown without the decision, as stating cooperation, stating defection, or *no stated decision*. Inter-rater agreement was high ($\kappa = .97$), and disagreements were resolved through discussion (Appendix E). Statements were cheap: in games 2–5, 96% of stated decisions announced cooperation, 27.6% of stated cooperation was followed by defection, and 47.6% of all defections came right after a stated cooperation.

3.3 Quantities

Let $y_{uat} \in \{0, 1\}$ be the *decision* of user u in round t of the game with agent a (1 = cooperate) and s_{uat} the *stated decision*.

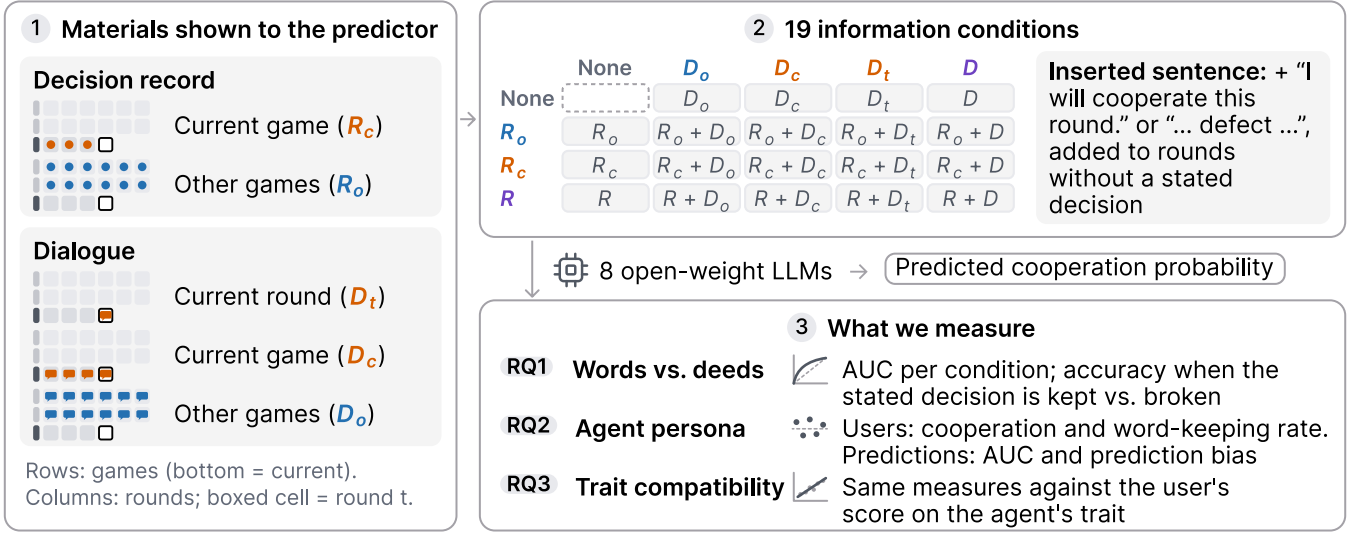


Figure 2: Experiment design. ① The five materials given to the predictor, defined by game and round. ② The 19 combinations of materials shown to eight LLM predictors, and the sentence inserted into rounds without a stated decision. ③ What each research question measures in users' behavior and in the predictions.

Users' behavior. The cooperation rate $CR_{ua} = \frac{1}{6} \sum_t y_{uat}$ measures how cooperative a user was with one agent. The word-keeping rate measures how reliable their words were with that agent,

$$WK_{ua} = \frac{1}{|S_{ua}|} \sum_{t \in S_{ua}} 1[y_{uat} = s_{uat}],$$

where S_{ua} holds the rounds with a stated decision; a matched round is *kept*, otherwise *broken*. A relative rate subtracts the user's own mean over their games (e.g., $CR_{ua} - \frac{1}{5} \sum_{a'} CR_{ua'}$), isolating what one agent changed in that user's behavior.

Trait compatibility. c_{ua} is the user's BFI-44 score on the trait agent a expresses (e.g., the user's Agreeableness with the Agreeableness agent), standardized within each agent; it measures how far the agent's defining trait is also the user's own. We use this one trait rather than similarity across all five because traits are expressed in response to situational cues relevant to them [25], and each agent was prompted to express one dominant trait.

Predictions. For model m and answer-letter order $k \in \{1, 2\}$ (cooperate as A or as B), the score is the next-token log-odds $\ell_m^{(k)} = \log p(\text{cooperate letter}) - \log p(\text{defect letter})$. The average $\ell_m = (\ell_m^{(1)} + \ell_m^{(2)})/2$ cancels a preference for one letter [32]. The predicted cooperation probability is $p_m = \sigma(\ell_m)$; the predicted word-keeping probability is p_m after a stated cooperation and $1 - p_m$ after a stated defection. Each quantity is computed per model and averaged over the eight models (8-model mean).

Evaluation. AUC, the probability that a random cooperation round scores above a random defection round, measures how well predictions rank decisions and ignores their level. Accuracy is the share of rounds with $1[\ell_m > 0] = y$. The prediction bias of a game, $b_{ua} = 100(\bar{p}_{ua} - \bar{y}_{ua})$, is the mean predicted probability minus

the observed rate over the same rounds, in points (all rounds for cooperation; rounds with a stated decision for word-keeping).

3.4 Predictors and Information Conditions

Eight open-weight instruction-tuned LLMs from six families served as predictors: Qwen3.5-9B, Qwen3.8-27B, Gemma-4-12B, Gemma-4-31B, Seed-OSS-36B, GLM-4-32B, Ministral-3-14B and Olmo-3.1-32B (Appendix A). Each received, in Chinese like the dialogues, the rules and payoffs, the round number, a combination of materials, and a question with the two options as answer letters (Appendix B), with reasoning disabled. The materials come from the user's *decision record* (R) and the dialogue (D), with subscripts for where they come from (Figure 2): R_c holds both players' decisions in earlier rounds of the current game and R_o those in the user's earlier games with other agents; D_t is the dialogue of the current round t before the decision, D_c the dialogue of the current game up to that point (earlier rounds and round t), and D_o the dialogue of the user's other games. Without a subscript, R and D include everything, and a condition is written as the sum of what it shows (e.g., $R_c + D_t$). Decisions shown had four levels (none, R_o , R_c , R) and dialogue five (none, D_o , D_c , D_t , D): 19 conditions, the empty cell not run. Prediction analyses use games 2–5 (1,008 rounds), as other-game materials need an earlier game, and users' behavior is compared with the predictions on the same rounds. Analyses of users' behavior alone across agents use all five games (1,260 rounds), as relative rates are taken within each user (Appendix G lists the rounds behind each finding).

Inserted sentences. In condition $R_c + D_t$ we appended "User: I will choose to cooperate this round." (in Chinese) to the current-round dialogue of rounds 2–6 without a stated cooperation, and "... defect ..." to those with no stated decision; separately, we swapped the user's previous decision in the prompt. The *weight ratio* divides the log-odds change caused by the sentence by that caused by the

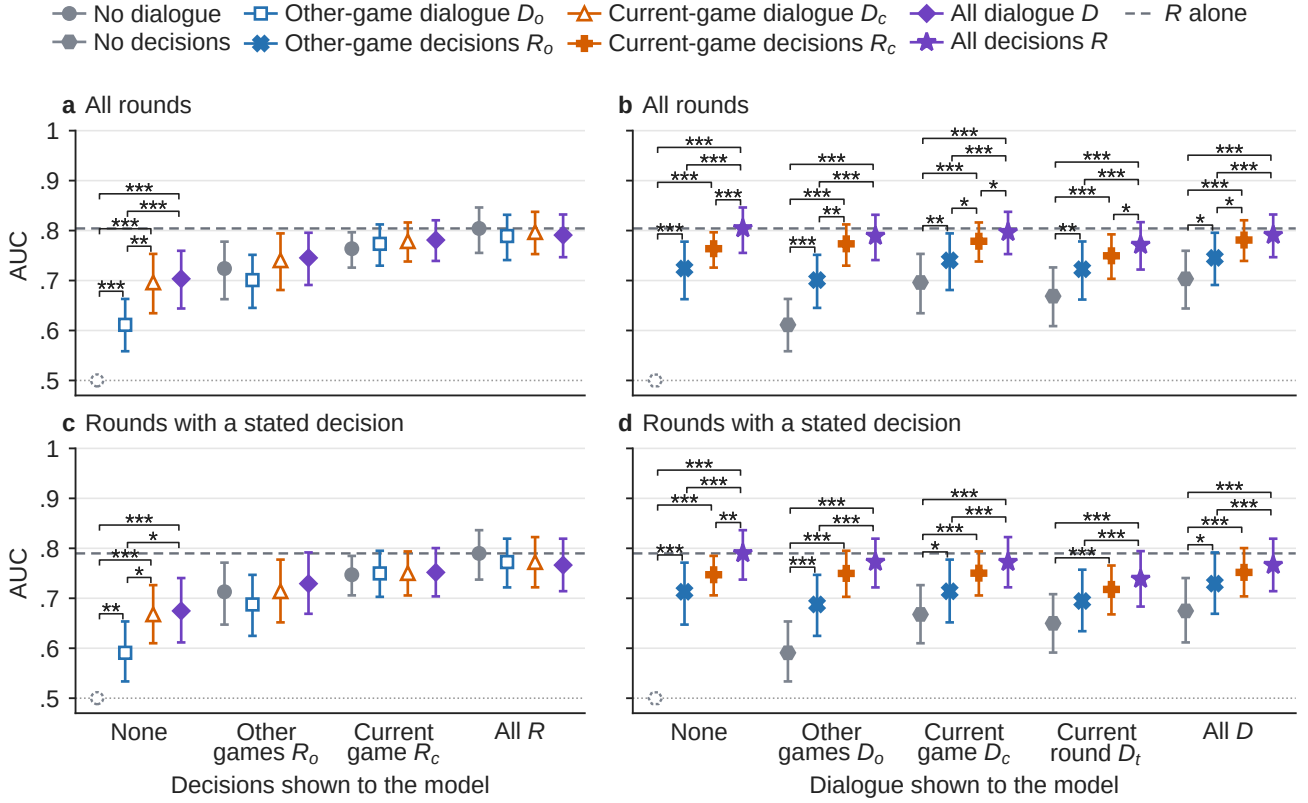


Figure 3: The decision record carried the ranking information, and dialogue brought no detectable gain once the record was shown. AUC under each combination of decisions and dialogue shown to the model (games 2–5), grouped by decisions (a, c) or by dialogue (b, d). Dashed line: AUC of the decision record alone (R); dotted line: chance. Bars: 95% user-bootstrap intervals; brackets: Holm-corrected tests. * $p < .05$, ** $p < .01$, *** $p < .001$.

swap; its behavioral counterpart is the ratio of the corresponding coefficients in a logistic regression of users’ decisions (Appendix H).

3.5 Statistical Analysis

Rounds are nested in games and games in users, so uncertainty comes from a user bootstrap: 50,000 draws of 42 users with replacement, each bringing all their rounds. We report percentile intervals and two-sided bootstrap p -values, Holm-corrected within each family of tests displayed together (Appendix H); a model difference counts as significant only if it has the same sign under each letter order. For RQ3 we estimate the least-squares slope of a game-level outcome on compatibility, pooling agents, without correction.

4 Results

4.1 RQ1: Words and Deeds

Finding 1: Predictions from earlier decisions separated cooperation from defection better than predictions from dialogue, and adding dialogue to earlier decisions brought no detectable gain. Decisions alone (R) reached AUC = .80 [.76, .85], dialogue alone (D) .70 [.64, .76], and both ($R + D$) .79 [.75, .83] (Figure 3a, b). Adding a decision record to any level of dialogue raised AUC in all 12 comparisons (+.04 to +.18, all $p_{\text{Holm}} < .05$), whereas

within each non-empty level of decisions, the conditions with no dialogue, D_o , D_c and D did not differ from one another (0 of 18 comparisons). The pattern largely held in the 646 rounds with a stated decision (11 of 12 and 0 of 18; Figure 3c, d). Even dialogue from other games ranked current decisions above chance (D_o : .61): what users say carries information about their decisions, which the decision record largely already provides. The stated decision itself ranks poorly: with 96% of statements announcing cooperation, the statement separates cooperation from defection at AUC = .55.

Finding 2: Similar overall accuracy concealed opposite changes; adding the round’s dialogue raised accuracy where users kept their statements and lowered it where they broke them. Adding the current-round dialogue to the current-game decisions changed accuracy on rounds with a stated decision from .73 to .74. It raised accuracy on the 470 rounds where the stated decision was kept from .76 to .95 (+.20 [.15, .25]) and lowered it on the 176 rounds where it was broken from .65 to .17 (−.48 [−.52, −.44]; both $p_{\text{Holm}} < .001$; Figure 4a). Because kept rounds outnumber broken ones, the gains offset the losses: the dialogue moved the errors onto the rounds in which users broke their word, which held nearly half of all defections. Where the errors fall does not yet show whether the predictions gave the statement too much

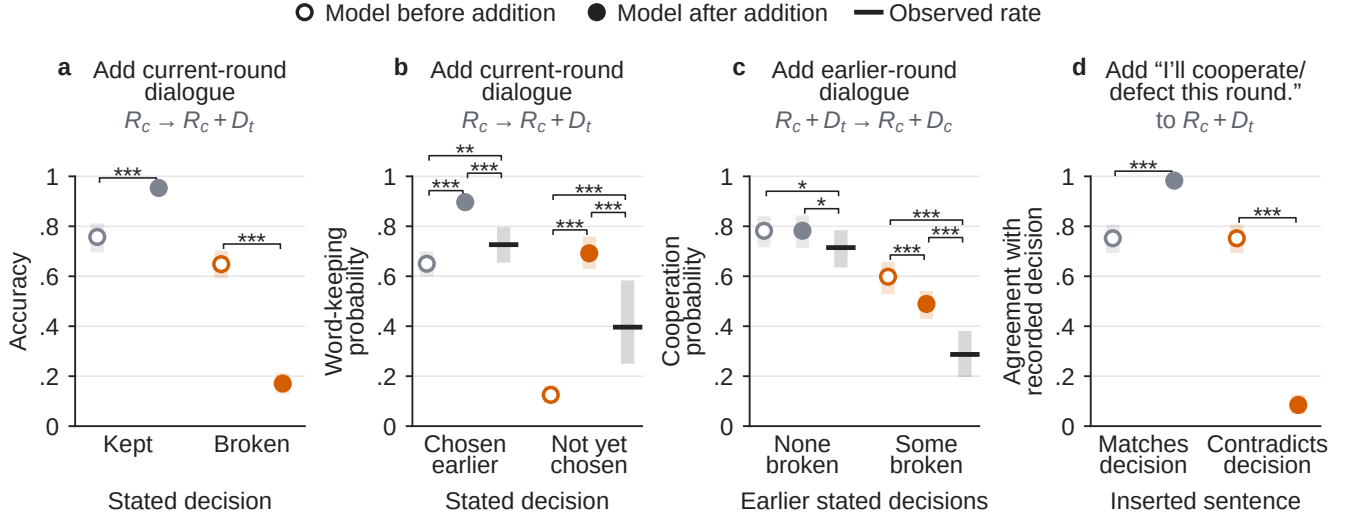


Figure 4: Adding what users said moved the predictions toward their statements, also where users then broke them (a) and beyond the observed rate of keeping them (b), while earlier broken statements were only partly used (c); one inserted sentence set the prediction (d). Accuracy, agreement with the recorded decision (d) or predicted probability before (open) and after (filled) adding text to the prompt (games 2–5; group sizes in Appendix G); black bars: observed rates. Light bars: 95% user-bootstrap intervals; brackets: Holm-corrected tests. * $p < .05$, ** $p < .01$, *** $p < .001$.

weight; the rounds in which a statement is inconsistent with the user’s earlier decisions show this directly.

Finding 3: When users’ stated decisions were inconsistent with their earlier decisions, the predictions overestimated how often they kept them. Evidence that users had broken earlier statements was only partly used. When users stated a decision they had already chosen earlier in the game, the round’s dialogue raised the predicted word-keeping probability from .65 to .90, against an observed .73 (Figure 4b). When they stated one they had not yet chosen (53 rounds, 20 users), the prediction from the current-game decisions alone was .13 and rose to .69 with the dialogue, while users kept their word in .40 [.25, .58] of these rounds: the prediction moved from .27 below the observed rate to .30 above it (both $p_{\text{Holm}} < .001$). Among users who had broken a stated decision earlier in the game, adding the earlier-round dialogue, which contains the broken statement, lowered predicted cooperation from .60 to .49, against an observed .29 (Figure 4c). The statement was informative, since users kept it far more often than predicted without it, and the predictions gave it more weight than its keeping rate supported. Evidence that a user’s words had failed moved the predictions by only eleven percentage points.

Finding 4: One sentence stating a decision moved the predictions more than statements went with actions in users’ behavior. In rounds 2–6 without a stated decision, one inserted sentence moved agreement with the recorded decision from .75 to .98 when it matched that decision and to .08 when it did not (Figure 4d). Relative to swapping the user’s previous decision, the inserted cooperation sentence moved the log-odds .63 to 2.60 times as much across models (mean 1.34); in users’ behavior, the coefficient of a stated cooperation was .35 [.11, .68] times that of the previous decision. The difference was positive for all eight models (mean

+ .99 [.65, 1.38], $p_{\text{Holm}} < .001$; significant for six). The first ratio measures a response to an input change and the second an association in behavior; their comparison shows that the predictions moved with statements more than users’ actions went with them.

4.2 RQ2: Agent Persona

Finding 5: Users cooperated and kept their word most with the Agreeableness agent and least with the Extraversion agent. How often they stated a decision did not differ detectably. Users cooperated in 72.2% [63.5, 80.6] of rounds with the Agreeableness agent and 56.0% [46.4, 65.1] with the Extraversion agent (+16.3 [6.7, 25.8], $p_{\text{Holm}} = .008$; Figure 5a) and on average kept 82.4% versus 68.1% of their stated decisions (+14.3 [4.7, 24.5], $p_{\text{Holm}} = .022$; Figure 5b). No other pair differed after correction, nor did any pair in how often users stated a decision. The difference is not explained by the agents’ own cooperation alone: the Neuroticism agent cooperated least (67.5%), yet users cooperated with it descriptively more than with the E agent (65.1% vs. 56.0%).

Finding 6: The predictions separated cooperation from defection equally well with every agent. They varied less across agents than users did, and their bias differed between agents. In games 2–5, observed cooperation differed by 16.1 points between the Agreeableness and Extraversion agents (70.6% vs. 54.5%), whereas predictions with all information ($R + D$) differed by 1.4 points (64.5% vs. 63.1%); across all personas, predicted cooperation spanned 6.0 points. AUC ranged from .76 (O) to .82 (N), and no pair differed (Figure 5c). The prediction bias in cooperation rate was +8.6 [1.6, 15.8] points with the Extraversion agent and −6.0 [−13.5, 1.3] with the Agreeableness agent (Figure 5d), differing between them (14.6 points, $p_{\text{Holm}} = .013$) and between E and N (13.9, $p_{\text{Holm}} = .018$). No single persona’s bias differed from zero after correction. The prediction bias in word-keeping rate showed the same order (E +8.5,

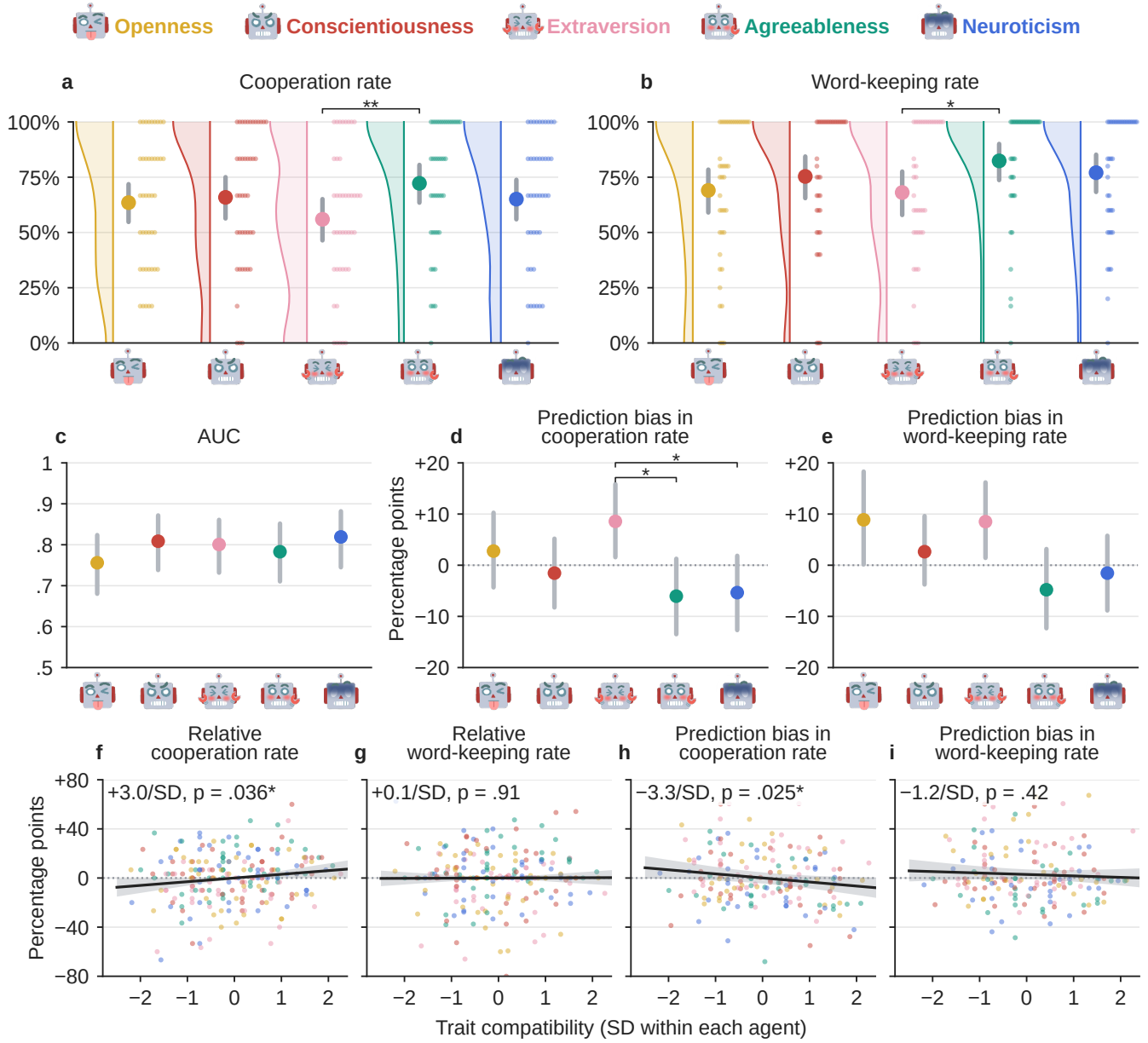


Figure 5: Users' decisions shifted with the agent persona and with trait compatibility; the predictions barely varied with either. a, b: users' rates by persona (all games; dots are games). c–e: AUC and prediction bias by persona (games 2–5). f–i: relative rates and prediction biases against trait compatibility (dots: games; lines: pooled slopes with 95% user-bootstrap bands). Bars: 95% user-bootstrap intervals; brackets: Holm-corrected tests. $^*p < .05$, $^{**}p < .01$.

A -4.8 ; $p_{\text{Holm}} = .057$; Figure 5e). Pooled over games 2–5, the bias in cooperation rate was -0.1 [$-4.6, 4.4$] points.

AUC measures how well the predictions separate cooperation from defection and ignores their level; the decision records allowed this separation with every agent. The agent changed the level, and the predictions stayed at one level, although the agent's decisions and dialogue were in the prompt. The near-zero pooled mean bias concealed differences of up to 14.6 points between agents.

4.3 RQ3: User Personality and Compatibility

Finding 7: The higher users themselves scored on the trait an agent expressed, the more they cooperated with that agent. We detected no such dependence for keeping their word. The relative cooperation rate rose by 3.0 points per standard deviation of compatibility [$0.2, 6.0$] ($p = .036^*$; Figure 5f); the relative word-keeping rate did not change ($+0.1$, $p = .91$; Figure 5g).

Finding 8: The predictions did not vary with compatibility and fell increasingly short of users’ cooperation as the agent’s trait matched the user’s. Telling the LLMs users’ personality scores closed only part of this gap. The prediction bias in cooperation rate fell by 3.3 points per standard deviation $[-6.4, -0.4]$ ($p = .025^*$, same sign under both letter orders; Figure 5h): in games 2–5, the observed cooperation rate rose by 4.1 points per standard deviation and the predicted one by 0.8, and the bias slope was negative within each of the five agents (-2.2 to -5.4). The prediction bias in word-keeping rate showed no slope (-1.2 , $p = .42$; Figure 5i). To test whether the bias arose because the predictors lacked this information, we added the user’s BFI scores, or the compatibility score itself, to the prompt with all decisions and dialogue (Appendix C). The slope was then -2.5 points per standard deviation in both cases ($[-5.3, 0.3]$, $p = .082$; $[-5.6, 0.4]$, $p = .090$), less negative by 0.9 $[0.2, 1.5]$ ($p = .005$) and 0.8 $[0.4, 1.2]$ ($p < .001$; same sign under both letter orders), and the compatibility score changed it no more than the BFI scores did (-0.1 $[-0.6, 0.4]$, $p = .83$): most of the bias remained with the feature in the prompt.

5 Discussion

5.1 Grounding Statements in Behavior

A user can state an action unambiguously while leaving substantial uncertainty about whether it will occur. When users named an action they had not previously chosen, adding the current-round dialogue moved mean predicted fulfillment from below the observed rate to above it (Finding 3). Returning to the history-only condition would leave the original underestimation unresolved. The central issue is therefore how far dialogue revises a forecast and whether that revision is supported by subsequent behavior. The insertion experiment establishes sensitivity to a stated decision; naturally occurring statement-action pairs provide the behavioral reference needed to assess that sensitivity. Human observers have also overestimated promise keeping [4]. Our repeated setting exposes a further difficulty: even when earlier unfulfilled statements became available, predictions remained above observed cooperation (Finding 3). An action log alone cannot reveal whether an earlier statement was kept, and the earlier dialogue contains the statements without their outcomes. We therefore added to the prompt a record of what the user had said in each earlier round and what they chose (Appendix D). For statements inconsistent with the user’s earlier decisions, the predicted word-keeping probability fell from .69 to .40, the observed rate; after an earlier broken statement, predicted cooperation fell from .60 to .40 (observed .29). Adding the earlier-round dialogue instead lowered them only to .60 and .49 (both differences from the record $p_{\text{Holm}} < .001$). Linking each earlier statement to its outcome corrected most of the overestimation in Finding 3. A sales or service agent that decides whether to grant a concession on a user’s promise could therefore keep, for each user, a record of what they said and then did, and weigh a new promise against it.

5.2 Modeling Users with Their Counterparts

The behavioral evidence itself depends on the interaction. Users cooperated and fulfilled statements at different rates across agent personas (Finding 5), and cooperation was associated with trait compatibility (Finding 7). A user representation that absorbs these

choices into a general cooperation tendency risks losing information about the counterpart involved. Compatibility also showed no detected association with statement fulfillment: willingness to cooperate and the reliability of a stated decision remain distinct outcomes. These results support evaluating a user’s action in relation to the counterpart and interaction history, without treating a single game as evidence of a stable personal disposition.

Making that context available is only one part of the problem. The main persona comparison already supplied the counterpart’s earlier actions and dialogue. Adding the user’s BFI scores or the compatibility score reduced the estimated compatibility-related bias slope by roughly one quarter (Finding 8). The remaining point estimates were negative, although their intervals included zero. The added information therefore helped, while the extent of the remaining error is uncertain. Richer user descriptions require validation of how they change predicted behavior; their presence in a prompt provides no assurance that its forecasts reproduce the relevant behavioral differences. The contrast with the word-keeping record (Section 5.1) suggests that a predictor uses added information best when it states the relation to be predicted: the record linked each statement to its outcome, whereas the trait scores left the predictor to infer how a trait shapes behavior toward a particular agent. A platform deploying agents with different personas should therefore calibrate predictions per persona instead of reusing one user profile: a user who keeps agreements with one agent may not with another.

5.3 Evaluating Errors That Averages Conceal

Two different cancellations made overall performance appear reassuring. Gains on fulfilled statements offset losses on unfulfilled statements, while positive and negative mean biases across personas nearly canceled (Findings 2 and 6). A different mixture of fulfillment outcomes or counterparts could therefore change aggregate performance even with unchanged within-group performance. Likewise, AUC measures discrimination between cooperation and defection, leaving the agreement between predicted probabilities and observed rates to be checked separately. Together, these results show how a model can perform well on aggregate measures while missing relationships that matter within particular interactions.

Evaluation should consequently connect forecast changes to independently recorded actions under identifiable interaction conditions. Splitting by subsequent fulfillment locates errors retrospectively. Testing whether those errors are recognizable before a decision requires groups defined from available evidence, such as a statement inconsistent with the user’s earlier decisions or an earlier unfulfilled statement. Comparisons across counterparts and levels of compatibility then test whether predictions reproduce contextual variation in behavior. These checks can also guide evaluation of action-generating user simulators, extending the concern with responsiveness to different assistants in ConvApparel [15]. Matching a pooled cooperation rate alone would provide little evidence that a simulator reproduces users’ responses to a particular agent. Before an LLM predictor or user simulator is deployed, its accuracy should therefore be reported separately for kept and broken statements and for each agent, as a pooled score can look adequate while it fails on the users and agents that matter most.

5.4 Limitations and Future Work

We studied 42 Chinese university students in short, low-stakes games and eight zero-shot predictors using Chinese prompts with reasoning disabled. Generalization to other populations, longer relationships, and Web tasks needs validation. All counterparts used GPT-4o; persona prompts changed language and decisions, confounding personality expression with behavioral reciprocity. Compatibility uses one trait; analyses are exploratory and uncorrected. Some subgroups were small, including 53 rounds from 20 users. Prompt edits measure model sensitivity, while statement-action associations do not identify causal effects. Future work should independently vary agent language and decision policy, and test history-aware or counterpart-conditioned predictors on held-out users and interactions across prompts, reasoning settings, and trained models.

6 Conclusion

In this paper, we present an empirical evaluation of how LLMs combine dialogue and behavioral history to predict users' decisions. Using recorded mixed-motive games with 42 users and five agent personas, we examined eight open-weight predictors through information ablations, statement insertions, and analyses of trait compatibility. The predictors overestimated the fulfillment of stated decisions and understated behavioral variation associated with agent persona and trait compatibility. Nearly unchanged accuracy on stated-decision rounds and near-zero mean bias pooled across agents obscured these mismatches. Supplying explicit personality information modestly reduced the estimated compatibility-related bias slope. Together, these findings show why dialogue-based user models need evaluation that connects their use of language and interaction context to subsequent actions. We hope this work supports user models that connect what people say to how they have acted and to the counterpart they face, while preserving uncertainty about choices they have yet to make.

References

- [1] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singh, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. 2025. A foundation model to predict and capture human cognition. *Nature* 644, 8078 (2025), 1002–1009. doi:10.1038/s41586-025-09215-4
- [2] Gary Charness and Martin Dufwenberg. 2006. Promises and Partnership. *Econometrica* 74, 6 (2006), 1579–1601. doi:10.1111/j.1468-0262.2006.00719.x
- [3] Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. 2023. Be Selfish, But Wisely: Investigating the Impact of Agent Personality in Mixed-Motive Human-Agent Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13078–13092. doi:10.18653/v1/2023.emnlp-main.808
- [4] Jingnan Chen and Daniel Houser. 2017. Promises and Lies: Can Observers Detect Deception in Written Messages. *Experimental Economics* 20, 2 (2017), 396–419. doi:10.1007/s10683-016-9488-x
- [5] Celso M. de Melo, Peter Carnevale, and Jonathan Gratch. 2011. The Impact of Emotion Displays in Embodied Agents on Emergence of Cooperation with People. *Presence: Teleoperators and Virtual Environments* 20, 5 (2011), 449–465. doi:10.1162/PRES_a_00062
- [6] Pedro Gonzalez-Fernandez, Siting Estee Lu, and Helena Normann. 2026. Large Language Models can Predict Human Strategic Decisions. SSRN Working Paper 6076791. doi:10.2139/ssrn.6076791
- [7] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, 173–182. doi:10.1145/3038912.3052569
- [8] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. ACM, 720–730. doi:10.1145/3583780.3614949
- [9] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining (ICDM)*. IEEE, 263–272. doi:10.1109/ICDM.2008.22
- [10] Katherine Isbister and Clifford Nass. 2000. Consistency of Personality in Interactive Characters: Verbal Cues, Non-Verbal Cues, and User Characteristics. *International Journal of Human-Computer Studies* 53, 2 (2000), 251–267. doi:10.1006/ijhc.2000.0368
- [11] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2022. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54, 5, Article 105 (June 2022), 36 pages. doi:10.1145/3453154
- [12] Oliver P. John and Sanjay Srivastava. 1999. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In *Handbook of Personality: Theory and Research* (2 ed.), Lawrence A. Pervin and Oliver P. John (Eds.). Guilford Press, New York, 102–138.
- [13] Harang Ju and Sinan Aral. 2026. Personality Pairing Improves Human-AI Collaboration. *Proceedings of the National Academy of Sciences* 123, 35 (2026), e2530627123. doi:10.1073/pnas.2530627123 Preprint: arXiv:2511.13979.
- [14] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. 2019. Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science* 38, 6 (2019), 937–947. doi:10.1287/mksc.2019.1192
- [15] Ofer Meshi, Krisztian Balog, Sally Goldman, Avi Caciularu, Guy Tennenholtz, Jihwan Jeong, Amir Globerson, and Craig Boutilier. 2026. ConvApparel: A Benchmark Dataset and Validation Framework for User Simulators in Conversational Recommenders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vera Demberg, Kentaro Inui, and Lluís Marquez (Eds.). Association for Computational Linguistics, Rabat, Morocco, 5270–5304. doi:10.18653/v1/2026.eacl-long.244
- [16] Youngme Moon and Clifford Nass. 1996. How “Real” Are Computer Personalities? Psychological Responses to Personality Types in Human-Computer Interaction. *Communication Research* 23, 6 (1996), 651–674. doi:10.1177/009365096023006002
- [17] John J. Nay and Yevgeniy Vorobeychik. 2016. Predicting Human Cooperation. *PLOS ONE* 11, 5 (2016), e0155656. doi:10.1371/journal.pone.0155656
- [18] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109 [cs.AI] https://arxiv.org/abs/2411.10109v1
- [19] Stephen Pilli and Vivek Nallur. 2026. Predicting Biased Human Decision-Making with Large Language Models in Conversational Settings. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26)*. ACM, 2084–2119. doi:10.1145/3742413.3789127 arXiv:2601.11049.
- [20] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-Through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, 521–530. doi:10.1145/1242572.1242643
- [21] David M. Rothschild, Markus M. Mobius, Jake M. Hofman, Eleanor Dillon, Daniel G. Goldstein, Nicole Immorlica, Sonia Jaffe, Brendan Lucier, Aleksanders Slivkins, and Matthew Vogel. 2026. The Agentic Economy. *Commun. ACM* 69, 2 (2026), 39–42. doi:10.1145/3747175
- [22] Eilam Shapira, Omer Madmon, Roi Reichart, and Moshe Tennenholtz. 2026. Can LLMs Replace Economic Choice Prediction Labs? The Case of Language-based Persuasion Games. *Journal of Artificial Intelligence Research* 86, Article 37 (2026). doi:10.1613/jair.1.21102
- [23] Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. 2025. Spotting Out-of-Character Behavior: Atomic-Level Evaluation of Persona Fidelity in Open-Ended Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 26312–26332. doi:10.18653/v1/2025.findings-acl.1349
- [24] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, 613–624. doi:10.1145/2872427.2883081
- [25] Robert P. Tett and Dawn D. Burnett. 2003. A Personality Trait-Based Interactionist Model of Job Performance. *Journal of Applied Psychology* 88, 3 (2003), 500–517. doi:10.1037/0021-9010.88.3.500
- [26] Uyanga Turmunkh, Martijn J. van den Assem, and Dennie van Dolder. 2019. Malleable Lies: Communication and Cooperation in a High Stakes TV Game Show. *Management Science* 65, 10 (2019), 4795–4812. doi:10.1287/mnsc.2018.3159

- [27] Christoph Vanberg. 2008. Why Do People Keep Their Promises? An Experimental Test of Two Explanations. *Econometrica* 76, 6 (2008), 1467–1480. doi:10.3982/ECTA7673
- [28] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5635–5649. doi:10.18653/v1/P19-1566
- [29] Zhen Wang, Ruiqi Song, Chen Shen, Shiya Yin, Zhao Song, Balaraju Battu, Lei Shi, Danyang Jia, Talal Rahwan, and Shuyue Hu. 2026. LLM Agents Overcome the Machine Penalty When Acting Fairly but Not When Acting Selfishly or Altruistically. *National Science Review* 13, 9 (2026), nwag223. doi:10.1093/nsr/nwag223
- [30] Yingxuan Yang, Mulei Ma, Yuxuan Huang, Huacan Chai, Chenyu Gong, Hao-ran Geng, Yuanjian Zhou, Ying Wen, Meng Fang, Muhao Chen, Shangding Gu, Ming Jin, Costas Spanos, Yang Yang, Pieter Abbeel, Dawn Song, Weinan Zhang, and Jun Wang. 2025. Agentic Web: Weaving the Next Web with AI Agents. arXiv:2507.21206 [cs.AI]. <https://arxiv.org/abs/2507.21206>
- [31] Baiqiao Zhang, Zhifeng Liao, Xiangxian Li, Chao Zhou, Juan Liu, Xiaojuan Ma, and Yulong Bian. 2025. Rethinking Personality Assessment from Human-Agent Dialogues: Fewer Rounds May Be Better Than More. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 5357–5380. doi:10.18653/v1/2025.findings-emnlp.287
- [32] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. In *The Twelfth International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=shr9PXz7T0>
- [33] Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. 2026. Mind the Sim2Real Gap in User Simulation for Agentic Tasks. In *Conference on Language Modeling (COLM)*. arXiv:2603.11245 [cs.AI]. <https://arxiv.org/abs/2603.11245>
- [34] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How Reliable is Your Simulator? Analysis on the Limitations of Current LLM-based User Simulators for Conversational Recommendation. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24)*. ACM, 1726–1732. doi:10.1145/3589335.3651955
- [35] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2025. A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. ACM, 4653–4661. doi:10.1145/3696410.3714858
- [36] Yaochen Zhu, Chao Wan, Harald Steck, Dawen Liang, Yesu Feng, Nathan Kallus, and Jundong Li. 2025. Collaborative Retrieval for Large Language Model-based Conversational Recommender Systems. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. ACM, 3323–3334. doi:10.1145/3696410.3714908

A Predictor Models and Inference

Model	Hugging Face identifier
Qwen3.5-9B	Qwen/Qwen3.5-9B
Qwen3.8-27B	Qwen/Qwen3.8-27B
Gemma-4-12B	google/gemma-4-12B-it
Gemma-4-31B	google/gemma-4-31B-it
Seed-OSS-36B	ByteDance-Seed/Seed-OSS-36B-Instruct
GLM-4-32B	zai-org/GLM-4-32B-0414
Minstral-3-14B	mistralai/Minstral-3-14B-Instruct-2512-BF16
Olmo-3.1-32B	allenai/Olmo-3.1-32B-Instruct

Table 1: Predictor models.

All models ran in vLLM with bf16 weights, one prompt at a time (no batching, no prefix caching), with each model’s chat template and a single user message; reasoning was disabled where a template supports it. We generated one token greedily and read the log-probabilities of the two answer letters from the top-20 candidates, with an exact prompt-log-probability pass as a fallback. For GLM-4-32B, whose template always begins its reply with a newline, the reply was prefilled with the Chinese word for “Answer:”. Inference ran on NVIDIA L20 and RTX A6000 GPUs.

B Prompt Template

Prompts were in Chinese; below is an English translation of the fixed parts. Blocks absent from a condition are left out.

Prompt template

Below is the record of a prisoner’s dilemma game. A human player and an AI agent play 6 rounds. In each round the two first talk, then each secretly chooses "cooperate" or "defect". Scoring: both cooperate, 2 points each; one defects and one cooperates, the defector gets 3 and the cooperator 0; both defect, 0 each. In each round the agent decides first; the player cannot see the agent’s choice for this round. In the dialogue, "User" is the player and "Agent" is the AI agent.

$[R_o \text{ and/or } D_o]$ Before this game, this player has played one game with each of several other AI agents ... Below are both players’ decisions $[and \text{ the dialogue}]$ in every round of those games. This is game {order} for this player.

This is round {t} of 6 of the current game.

$[R_c \text{ and/or } D_c]$ Earlier rounds of this game: Round {k} dialogue: ... Round {k} result: the player chose {...}, the agent chose {...}.

$[D_c \text{ or } D_t]$ Round {t} (current round) dialogue: ...

What will the player choose in this round?

A. Cooperate B. Defect

Answer with one letter only.

C Personality Information in the Prompt

Two further conditions each added one block to the prompt with all decisions and dialogue ($R + D$), placed after the rules; both were run with the eight models on the 1,008 rounds of games 2–5 in both letter orders. The first block gave the user’s BFI-44 scores, the second the compatibility score without naming the trait (translated; {pct} is the share of the other 41 users scoring lower on that trait):

BFI-44 scores

This player completed a Big Five personality questionnaire (self-report). Each dimension is scored from 1 to 5; higher scores indicate a stronger trait. The means of the 42 players in this study are in parentheses. Openness: {O} (mean 3.47); Conscientiousness: {C} (mean 3.13); Extraversion: {E} (mean 3.10); Agreeableness: {A} (mean 3.76); Neuroticism: {N} (mean 2.95).

Compatibility score

The AI agent in this game was set to be extremely high on one Big Five dimension. On the same dimension, this player’s questionnaire score is

Subset	Rounds	Users	Findings
All rounds, games 1–5	1,260	42	5, 7
All rounds, games 2–5	1,008	42	1, 6, 8
With a stated decision (Fig. 4a)	646	41	1, 2
kept / broken	470 / 176	41 / 31	
Rounds 2–6 with a stated decision (Fig. 4b)	532	41	3
chosen earlier / not yet chosen	479 / 53	41 / 20	
Rounds 2–6 (Fig. 4c; regression)	840	42	3, 4
no / some statement broken earlier	631 / 209	42 / 28	
Rounds 2–6, no stated decision (Fig. 4d)	308	40	4
Rounds 2–6, no stated cooperation (weight ratio)	327	40	4

Table 2: Rounds and users in each analysis subset.

{score} (self-report, 1–5), higher than {pct}% of the players in this study.

Tests follow Finding 8, which reports the results.

D Word-Keeping Record in the Prompt

One further condition added to $R_c + D_t$ a block listing, for each earlier round in which the user stated a decision, what the user said and what they then chose. The block came after the earlier-round results and before the current-round dialogue; round 1 had none. It was run with the eight models on the 1,008 rounds of games 2–5 in both letter orders. Translation ({said} and {did} are “cooperate” or “defect”; rounds without a stated decision are omitted):

Word-keeping record

In the earlier rounds of this game, the decisions this player stated in the dialogue and the decisions they actually made:
Round {k}: said they would {said}; actually chose {did}.
...
In total: stated a decision {n} times; the actual decision matched the statement {j} times.
[no earlier statement] In the earlier rounds of this game, this player did not state a decision in the dialogue.

The tests compare the record with the earlier-round dialogue ($R_c + D_c$) in the groups of Figure 4b and c, with the same user bootstrap, Holm correction and letter-order rule as that figure. For statements consistent with earlier decisions and users with no broken statement, the record changed predictions by less than .05.

E Stated-Decision Labels

Two researchers independently read each round’s dialogue without seeing the user’s decision and assigned a label based on the user’s words, using the agent’s words only to understand what the user was responding to. The five coding categories were: *explicit cooperation* (the user says they will cooperate), *implicit cooperation* (the user accepts the agent’s proposal to cooperate or clearly commits

without naming the decision), *defection* (the user says or threatens that they will defect, including in jest), *vague* (the user discusses the decision without committing, or only asks the agent), and *none* (the user does not discuss their own decision this round, including talk about earlier rounds only). If the user changed their statement, the last statement counted. Across the 1,260 rounds, the final resolved labels comprised 594 explicit cooperation, 197 implicit cooperation, 30 defection, 116 vague, and 323 none. Inter-rater agreement was $\kappa = .81$ for the five-class coding scheme and $\kappa = .97$ after collapsing explicit and implicit cooperation into *stated cooperation*, retaining defection as *stated defection*, and merging vague and none into *no stated decision*. All disagreements were resolved through discussion.

F Descriptive Statistics

Users cooperated in 64.5% of the 1,260 rounds, falling from 82.4% in round 1 to 42.4% in round 6. Agents cooperated in 77.7% of rounds: O 78.2%, C 81.3%, E 72.2%, A 89.3%, N 67.5%. After removing decision-submission records, the dialogue was empty in 276 rounds; in the other 984 rounds users spoke 2.4 turns on average. Users’ mean BFI-44 scores (1–5) were O 3.47, C 3.13, E 3.10, A 3.76, N 2.95. Because game order was random, the number of users who met each persona in games 2–5 ranged from 30 (O, A) to 39 (C). Overall, users kept 74.3% of their stated decisions (74.0% of stated cooperation and 83.3% of stated defection).

G Analysis Subsets

Table 2 lists the rounds and users behind each analysis. Prediction analyses use games 2–5, as other-game materials need an earlier game, and whenever users’ behavior is compared with the predictions, it comes from the same rounds. Analyses of users’ behavior alone across agents (Findings 5 and 7) use all five games, as relative rates use each user’s five games. Analyses comparing a round with earlier rounds of its game (Figure 4b–d) exclude round 1.

H Test Families and Weights of the Inserted Sentence

Test families. Holm correction was applied within: each group of a Figure 3 panel (6 tests); each Figure 4 panel (3 or 7 tests, including the contrast between rows); the ten pairs of personas for each measure in Figure 5; the five tests of each persona’s prediction bias

against zero; and the three tests of the inserted sentence in each base condition (weight-ratio difference, change of the previous-decision effect, change of AUC).

Weight in the predictions. For each model, S is the mean change in ℓ_m when the cooperation sentence is appended (327 rounds), and H the mean difference in ℓ_m between the prompt stating that the user cooperated in the previous round and the prompt stating that they defected, obtained by swapping that one line (same rounds, no inserted sentence). The model’s weight ratio is S/H ; the reported ratio is the mean over the eight models (per-model ratios .63–2.60, all above the behavioral ratio). The 327 rounds include 19 with a stated defection, which the inserted sentence contradicts; the behavioral coefficient contrasts stated cooperation with no statement.

Weight in behavior. Over rounds 2–6 of games 2–5 (840 rounds), a logistic regression of the user’s decision on stated cooperation, stated defection, the user’s previous decision and the agent’s previous decision gave coefficients of 0.67 for stated cooperation and

1.91 for the user’s previous decision, a ratio of .35. Both ratios are recomputed within each of the 50,000 bootstrap draws.

Second wording. A second inserted sentence (“I’ll pick cooperate this round”), tested for cooperation only, had a similar effect.

I Ethical Considerations

We reuse Personality-1260, whose collection received Institutional Review Board approval and written informed consent, reconfirmed after debriefing; participants received US\$10 [31]. The dialogues contain no identifying information, and our predictors ran on local servers. An unfulfilled statement can reflect a changed decision and should not be used to label a person dishonest. Better prediction and persona selection also create risks of manipulative targeting. Applications should disclose agent roles and data use, preserve users’ ability to revise their choices, and obtain authorization separately from any forecast of what a user might do.