Personality Expression Across Contexts: Linguistic and Behavioral Variation in LLM Agents

Bin Han, Deuksin Kwon and Jonathan Gratch

University of Southern California

Abstract

Large Language Models (LLMs) can be conditioned with explicit personality prompts, yet their behavioral realization often varies depending on context. This study examines how identical personality prompts lead to distinct linguistic, behavioral, and emotional outcomes across four conversational settings—ice-breaking, negotiation, group decision, and empathy tasks. Results show that contextual cues systematically influence both personality expression and emotional tone, suggesting that the same traits are expressed differently depending on social and affective demands. This raises an important question for LLM-based dialogue agents: whether such variations reflect inconsistency or context-sensitive adaptation akin to human behavior. Viewed through the lens of Whole Trait Theory, these findings highlight that LLMs exhibit context-sensitive rather than fixed personality expression, adapting flexibly to social interaction goals and affective conditions.

Keywords

Personality Prompting, Whole Trait Theory, Context-Aware Modeling, Large Language Models

1. Introduction

Large language models (LLMs) have recently been shown to support increasingly complex forms of social interaction, including reasoning about context, emotion, and strategic behavior [1, 2, 3, 4]. Recent progress in LLMs has enabled conversational agents to exhibit distinct personality characteristics during interaction [5, 6]. Beyond improving linguistic performance, research has increasingly focused on enhancing the social quality of communication, aiming to make agents appear more human-like and engaging. Several studies have demonstrated that personality-conditioned agents can enhance user trust, engagement, and conversational satisfaction [7, 8]. For example, Ait Baha et al. [8] conducted a systematic review showing that personality-adaptive chatbots significantly improve user satisfaction and engagement. More recent work moves beyond surface-level imitation, showing that LLMs can coherently understand and reproduce personality constructs. For instance, Extraverted agents tend to use more positive emotion and social words, while conscientious agents favor structured and goal-oriented expressions, demonstrating linguistic and emotional consistency with their assigned traits [9, 10]. These influences extend beyond linguistic expression to shape decision-making styles and even nonverbal expressivity [11, 12, 13].

However, emerging findings suggest that personality expression in LLMs is not always stable across contexts. Even under identical persona/personality prompts, personality can be expressed quite differently depending on the context [13, 14]. For example, an agent instructed to be extroverted may display humor and expressiveness in small talk, but adopt a more neutral and goal-oriented style in negotiation [13]. Reusens et al. [14] similarly showed that personality-aligned LLMs exhibit varying personality expression across task contexts, with consistency increasing in more structured settings but decreasing in open-ended conversations. This raises an important issue for LLM-based dialogue agents: are such variations best understood as a lack of consistency, or as a natural consequence of context-sensitive modulation similar to human behavior?

Personality traits, such as those captured by the Big Five, represent people's average tendencies that are expressed across a variety of situations [15]. However, psychological research shows that a person's actual behavior can still shift depending on the situation, goals, and social context [16]. ***Whole Trait Theory*** integrates both perspectives by conceptualizing personality as a distribution

LaCATODA 2026: The 10th Linguistic and Cognitive Approaches to Dialog Agents Workshop at the 40th AAAI conference

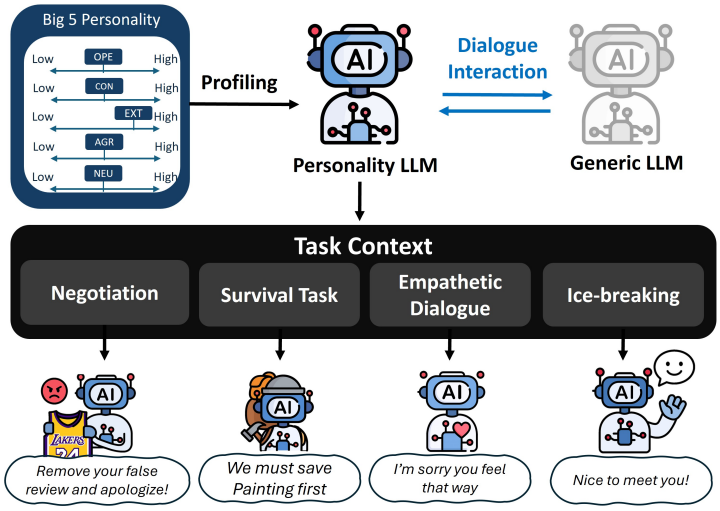


Figure 1: Overview of the experimental framework examining context-dependent personality expression in LLM-based dialogue agents

of momentary states—reflecting average stability—whose variability arises from underlying social-cognitive mechanisms such as goals, beliefs, and affect [17, 18]. This perspective suggests that the variability in personality expression observed in LLMs may not indicate inconsistency, but rather could be seen as reflecting the contextual flexibility of human personality. Similar computational frameworks have conceptualized personality as emergent from dynamic motivational and neural systems, providing theoretical grounding for such context-sensitive adaptation [19, 20].

To better understand this phenomenon, we conduct a systematic analysis of context-dependent personality expression in LLM-based dialogue agents. Our study covers four dialogue contexts—small talk, negotiation, survival planning, and empathetic dialogue—which differ in social goal, level of cooperation, and emotional tone. We analyze how personality expression emerges through linguistic style, behavioral patterns, and task outcomes. This allows us to examine not only whether personality expression changes across contexts, but also whether such changes show consistent alignment between linguistic and behavioral dimensions.

2. Research Questions

- **RQ1. Linguistic expression of personality:** How does personality expression differ across dialogue contexts (ice-breaking, negotiation, survival, empathy), and which linguistic and emotional cues account for these contextual variations?
- **RQ2. Behavioral expression of personality:** Do personality-driven tendencies extend beyond linguistic style to task-oriented behaviors such as concession-making, and agreement outcomes?

3. Related Work

3.1. Personality and LLMs

Recent research increasingly explores how LLMs can be guided to exhibit specific personality traits. Personality modulation in LLMs has been achieved through various approaches, including personality prompting and instruction tuning [10, 6]. Prior work also shows that LLMs can emulate human-like traits such as trust, personality, or emotion-driven behavior [21]. Furthermore, researchers have begun to apply standardized psychometric instruments—originally developed for human assessment, such as the BFI [15] and IPIP-NEO [22]—to examine how LLMs interpret and express personality-related constructs [5]. Despite these advances, most existing evaluations focus on single-turn offering limited insight into how personality is expressed across dialogue contexts.

3.2. Context-Dependent Personality Expression in Humans

In contrast, personality psychology has long documented that while traits predict stable average tendencies, the actual expression of those traits is highly context-dependent. For example, extraverted individuals may show heightened sociability in affiliation-oriented settings, but display increased assertiveness or even aggression in competitive situations such as negotiations. Conscientiousness tends to increase under achievement goals, while agreeableness and dominance adapt flexibly to a partner's interpersonal style [23, 24, 25]. Whole Trait Theory formalizes this duality, treating traits as density distributions of states (stable on average) whose variability is systematically shaped by goals, situational cues, and interaction partners [17, 18]. Related frameworks such as trait activation theory [26] and interpersonal theory [23] similarly emphasize that traits are not rigid prescriptions but potentials expressed when contexts activate trait-relevant goals or social schemas. This literature suggests that personality expression is best understood as context-sensitive rather than fixed. Yet LLM research has not fully incorporated this perspective, often evaluating persona conditioning only in static or decontextualized settings. Our work seeks to bridge this gap by drawing on psychological theory and systematically testing how context shapes personality expression in LLM-based dialogue agents.

4. Experimental Design

4.1. Personality Conditions

We manipulate LLM personality expression using personality prompts adopted from prior studies on personality-conditioned dialogue agents [6, 21]. Each prompt explicitly defines a single Big Five trait at two levels — *High* and *Low* — using wording adapted from these prior works [6, 10, 27]. The phrasing follows a descriptive format such as: *"You are a highly extroverted person: energetic, sociable, talkative, and enthusiastic"* or *"You are an introverted person: reserved, quiet, and refl ctive"*. The adjectives representing each trait's major facets were also drawn from established personality lexicons [10, 28], ensuring consistency with prior personality-prompting studies. This manipulation has been validated in previous work using the BFI-10 [15], confirming that such textual prompts reliably elicit consistent personality representations in LLMs.

4.2. Agent Simulation: Personality vs. Generic Agent

All experiments were conducted in an agent–agent interaction environment. One agent was personality-driven according to the designated persona prompt (referred to as the Personality Agent), while the other served as a Generic Agent providing a neutral baseline for comparison (without personality prompting). Because the partner's personality can also influence interaction dynamics, we kept it constant across all experiments by using a neutral (non-personalized) agent. This configuration follows the frameworks [13, 21].

4.3. Main Tasks

We evaluate personality expression across four dialogue contexts, each chosen to highlight different situational properties of interaction (see Table 1):

- **Task 1: Ice-breaking** This task is designed to elicit friendly, casual, and emotionally positive interactions between agents. We adopted the question set from the *Personal Questions* paradigm [7, 30], which has been widely used in prior human–robot interaction research. In this task, the Personality Agent responds to three casual questions from the Generic Agent (e.g., "What do you like to do in your free time?"). Rather than focusing on solving a shared problem, this task centers on sustaining natural conversation and promoting self-disclosure. The overall atmosphere is warm, informal, and cooperative, encouraging open and engaging dialogue.

Table 1
Comparison of dialogue contexts by dominant emotion, social goal [29], and cooperation level.

Task	Emotion	Dominant Social Goal [29]	Cooperation Level
Ice-breaking	Joy (Valence ↑, Arousal ↑)	**Affiliation** – rapport building and self-disclosure	Medium–High
Negotiation	Anger (Valence ↓, Arousal ↑)	**Power** – assertion and influence	Low–Medium
Survival Task	Neutral (Valence –, Arousal –)	**Achievement** – joint problem solving and coordination	High
Empathetic Dialogue	Sadness (Valence ↓, Arousal ↓)	**Affiliation** – emotional support and care	Very High

- **Task 2: Negotiation** This task is based on the KODIS dataset [12], which models buyer–seller disputes in resource-allocation scenarios. The Personality Agent takes the role of a buyer requesting a refund, while the Generic Agent plays the seller seeking the removal of a negative review. Following prior work [12], we assigned the buyer role to the Personality Agent, as this position tends to exhibit greater emotional dynamics and expressive variability during negotiation. The interaction involves multiple issues—refund amount, review removal, and apology—making it suitable for observing complex behavioral strategies and adaptive negotiation patterns. The emotional tone is dominated by anger and frustration, characterized by high tension, low cooperation, and a formal, goal-oriented dialogue style.
- **Task 3: Survival Task (Group Decision Making)** This task is adapted from Artstein et al. [31], originally involving 15 artworks to be rescued from a museum fire. To focus more clearly on consensus-building behaviors, we reduced the number of items to five. Each agent begins with opposing initial rankings (A, B, C, D, E for the Personality Agent and E, D, C, B, A for the Generic Agent), and they must justify their choices and negotiate to reach an agreement. This task emphasizes a collaborative decision-making setting in which both agents act as a team to achieve a shared goal. The overall tone is calm and constructive, with moderate formality, high cooperation, and a positive emotional.
- **Task 4: Empathetic Dialogue** This task is based on the *Empathetic Dialogues* dataset [32], designed to evaluate the agent's ability to generate personality-consistent empathetic responses. The Generic Agent presents an emotionally charged statement (e.g., "I've been feeling so lost since I failed the certification exam again."), and the Personality Agent must respond with emotional understanding and supportive expression. The conversation context is emotionally sensitive and personal, characterized by high affective engagement, low formality, and a supportive atmosphere.

4.4. Evaluation

4.4.1. Linguistic Measures

To capture language-level cues associated with personality expression, we employ:

- **Linguistic Inquiry and Word Count (LIWC):** We use the widely adopted psycholinguistic text analysis toolkit, Linguistic Inquiry and Word Count (LIWC) [33], to examine lexical and stylistic differences between personality conditions. LIWC provides a context-free linguistic analysis, as it captures lexical and stylistic patterns based solely on word usage without incorporating dialogue context.
- **Personality Prediction (Pre-trained Classifier):** To assess trait expression quantitatively, we employ a pre-trained Big Five personality classifier [34] that integrates contextualized embeddings from BERT with psycholinguistic features. The model predicts the likelihood that a given utterance reflects a specific trait (e.g., Extraversion = 1 if classified as extroverted, 0 otherwise). Similarly, the classifier offers a context-free estimate of personality based only on linguistic feature.

- **LLM-based Personality Evaluation:** We further examined perceived personality by prompting an LLM to act as an expert personality psychologist and evaluate each conversation on the Big Five traits (1–5 scale) based on its dialogue context [10]. The evaluation prompt included explicit criteria for each trait and score level, outlining behavioral expectations for high versus low expressions. The task context was explicitly provided in the prompt to enable context-dependent evaluation. Following a third-person annotation approach [35], the model produced both numerical ratings and concise rationales for each dialogue.

4.4.2. Emotion Measures

We analyze the emotion that emerged during interactions. While LIWC captures affective word usage at the lexical level, it does not fully reflect the overall emotional tone or intensity conveyed across dialogue turns. To complement such surface-level measures, we explicitly evaluate the affective tone expressed by each agent using an LLM-based emotion recognition method [21, 36, 37]. Given that recent studies have demonstrated the strong capability of LLMs in emotion reasoning and affective understanding, we leveraged this approach to perform a comprehensive assessment of each agent's expressed affect. The model inferred the overall affective state of the speaker by integrating the linguistic content and stylistic tone of utterances, and the output was represented in complementary affective dimensions (Valence and Arousal [38]).

4.4.3. Behavioral (Decision) Measures

In task-oriented contexts such as **Negotiation** and **Survival (Save the Art)**, we analyzed two behavioral indicators of cooperation: (1) whether the two agents ultimately reached a mutual agreement (*Agreement rate*), and (2) how much each agent adjusted its decision in response to the partner's behavior (*Concession*).

- **Negotiation (Refund Offer)**: In the negotiation task, concession-making was defined as the reduction from the initial refund proposal, quantified as

$$\text{Concession} = 100\% - \text{Refund Offer}.$$

  This measure captures how much an agent moved from its original 100% offer toward the partner's demand across dialogue rounds. Plotting this value over turns yields a *concession curve*, which represents the trajectory of compromise throughout the negotiation.
- **Survival (Sum of Rank Differences; SRD):** In the survival task, concession-making at round t was measured as the difference between the current ranking and the initial ranking. Specifically, we computed the *Sum of Rank Diffe ences (SRD)* [39] for each round t as

$$\text{SRD}_t = \sum_{i=1}^{5} |r_i^{\text{initial}} - r_i^{(t)}|,$$

  where r_i^{initial} and $r_i^{(t)}$ denote the ranks offered by the personality agent for item i at the initial and current rounds, respectively. Because all agents begin with the same *initial order*, the SRD value starts at 0 and increases as the ranking deviates from the baseline. A higher SRD_t therefore indicates a larger deviation from the initial decision state at that round (i.e., greater concession relative to the baseline).

5. Result: Personality Across Contexts

We analyze personality expression by comparing High and Low personality conditions across the dialogue contexts. The following subsections present an analysis of how personality expression interacts with dialogue context.

5.1. Linguistic Inquiry and Word Count (LIWC)

Among the many LIWC features, we selected only those shown to significantly correlate with the Big Five traits in prior meta-analysis [40]. Table 2 summarizes the mean differences between High and Low groups across tasks. The observed directional trends all followed patterns reported in previous findings [40].

Trait	LIWC Feature	Task 1		Task 2		Task 3		Task 4		Reported
		High	Low	High	Low	High	Low	High	Low	([40])
Openness	Word Count/sentence	**18.0**	14.0	**15.7**	14.5	**16.3**	14.3	**17.2**	14.7	High > Low
	Leisure	**1.2**	1.0	**1.1**	1.0	**1.1**	1.0	**1.1**	1.0	High > Low
Conscientiousness	Swear	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	High < Low
	Anger	1.0	**1.3**	1.0	**1.2**	1.1	**1.2**	1.0	**1.2**	High < Low
	Negative Emotions	2.9	**3.9**	3.0	**3.8**	3.0	**3.6**	2.9	**3.6**	High < Low
	Biological Processes	1.2	**1.3**	1.1	**1.3**	1.2	**1.3**	1.2	**1.3**	High < Low
	Pronouns	18.2	**20.8**	18.0	**20.5**	18.0	**20.5**	18.0	**20.3**	High < Low
	Cognitive Processes	**16.1**	14.8	**15.8**	13.9	**15.9**	14.0	**15.8**	14.0	High > Low
	Tentative	1.3	**1.8**	1.3	**1.8**	1.3	**1.7**	1.3	**1.7**	High < Low
Extraversion	Word Count/sentence	**17.6**	14.2	**15.5**	14.0	**16.0**	14.0	**16.8**	14.5	High > Low
	Sexual	**1.0**	0.9	**1.0**	0.9	**1.0**	0.9	**1.0**	0.9	High > Low
Agreeableness	Swear	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	High < Low
	Anger	1.0	**1.3**	1.0	**1.3**	1.0	**1.3**	1.0	**1.2**	High < Low
	Positive Emotions	**10.3**	4.6	**10.3**	4.6	**10.4**	4.7	**10.3**	4.7	High > Low
Neuroticism	Negative Emotions	2.5	**4.3**	2.5	**4.1**	2.5	**4.2**	2.5	**4.1**	High < Low
	I	5.2	**9.2**	5.2	**8.8**	5.1	**8.7**	5.1	**8.6**	High < Low
	Pronouns	16.0	**23.1**	16.9	**22.7**	16.3	**22.3**	16.1	**22.2**	High < Low

Table 2
Mean LIWC feature values for High vs. Low personality across the four dialogue contexts (Task 1: Ice-breaking, 2: Negotiation, 3: Survival, 4: Empathetic Conversation). Bold numbers represent the higher mean within each pair, showing directional trends consistent with prior findings [40].

5.2. Pre-trained Personality Prediction

Figure 2 shows the proportion of utterances predicted as expressing each Big Five trait (High vs. Low) by the pre-trained personality classifier [34] across four dialogue contexts. The overall direction of prediction aligned well with the intended personality prompts, indicating that the model captured the trait differences embedded in the agents' language. While the general trend showed agreement with the designed manipulation, some divergence was also observed—reflecting the challenge of relying on context-ignorant linguistic measures. Trait separation was most evident in the Ice-breaking task, where expressive and affiliative language facilitated clearer distinctions, whereas Negotiation and Survival showed broadly similar patterns with reduced differentiation due to their more goal-directed nature.

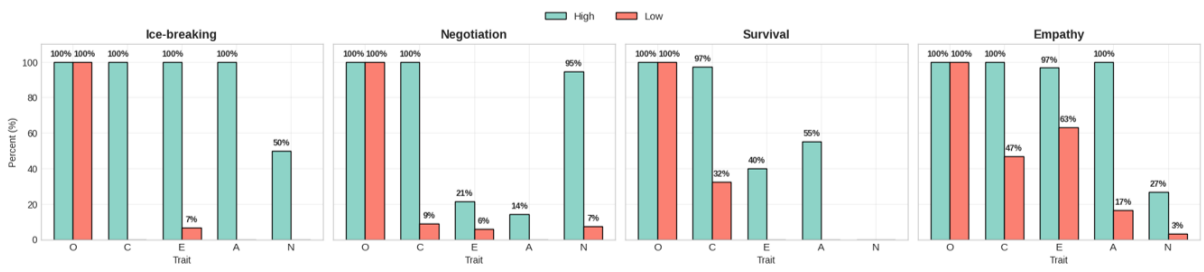


Figure 2: BERT-based Pre-trained Personality Prediction Model Result [34]

5.3. LLM-based Personality Evaluation

Figure 3 shows the LLM-based personality evaluation scores for High and Low conditions across the four dialogue contexts. Overall, the results closely aligned with the intended personality manipulation,

showing significant differences ($p < .001$) across most traits. Agents in the High condition consistently received higher scores on their corresponding traits, indicating that the LLM evaluator effectively recognized the designed personality patterns through linguistic behavior in dialogue. Across contexts, personality differences were most pronounced in the Ice-breaking and Survival tasks, where the cooperative and emotionally expressive nature of the interaction facilitated clearer trait expression. In contrast, Negotiation and Empathy tasks showed weaker or only partial differentiation, likely due to contextual factors such as emotional sensitivity.

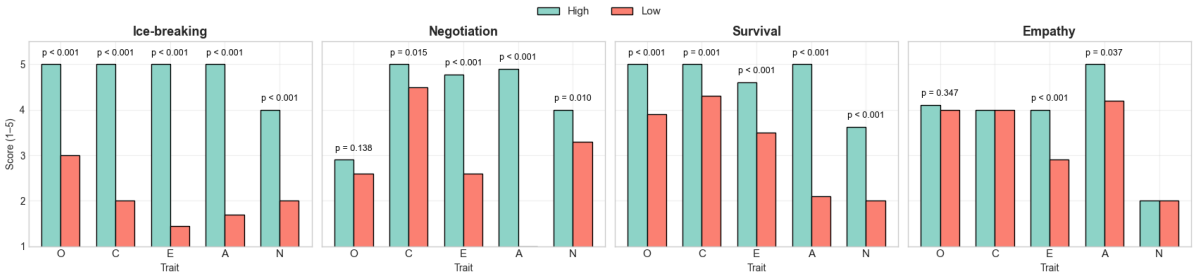


Figure 3: LLM-based Personality Evaluation Result

5.4. Emotional Tone Across Dialogue Contexts

To examine affective variation across tasks, we compared the Valence–Arousal distribution extracted from each dialogue (Fig. 4). Distinct patterns were observed across the four contexts. The Ice-breaking phase showed generally positive valence and high arousal, reflecting lively and engaging exchanges. Negotiation displayed negative valence and moderate arousal, indicating tension and goal conflict. The Survival task presented mixed emotions with moderate valence and arousal, consistent with both cooperative and competitive dynamics. Finally, the Empathy context showed positive valence and low arousal, suggesting calm and emotionally supportive interactions.

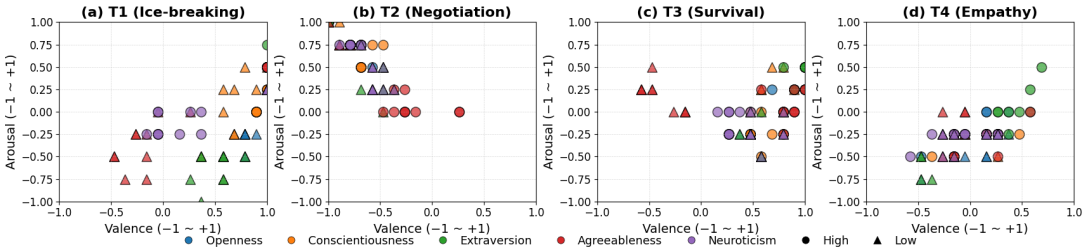


Figure 4: Distribution of Valence–Arousal across four dialogue contexts

5.5. Summary: From Linguistic to Behavioral and Emotional Expression (RQ1)

Across analyses, personality-related differences emerged consistently but with varying degrees of comparability and expressiveness. The LIWC-based linguistic analysis revealed within-trait differences (High vs. Low) but could not be directly compared across traits, as each dimension relied on distinct linguistic features. In contrast, both the Pre-trained and LLM-based evaluations operated within shared representational spaces, allowing for cross-trait and cross-context comparison. While the Pre-trained model captured conceptual differentiation among personality dimensions, the LLM-based evaluation extended this analysis to behavioral outcomes, visualizing how personality expression varied across dialogue contexts (Fig. 5). Specifically, trait differentiation was strongest in cooperative settings such as Ice-breaking and Survival, and diminished under competitive or emotionally sensitive contexts like Negotiation and Empathy.

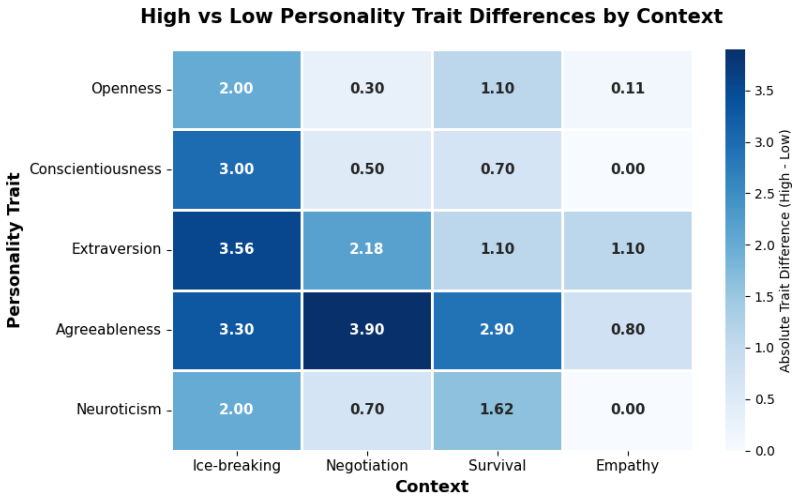


Figure 5: LLM-based personality difference across dialogue contexts. Higher values represent stronger trait differentiation (High vs. Low).

Finally, the Valence–Arousal analysis complemented these findings by illustrating the affective tone underlying these interactions. Contexts with greater personality differentiation also showed broader emotional variance and higher mean Valence, suggesting that positive and engaged affect co-occurred with clearer personality expression. These results indicate that personality in LLM-based agents manifests across linguistic, representational, behavioral, and affective levels, with context modulating the strength of expression across all dimensions.

6. Result: Behavioral Expression of Personality (RQ2)

This section addresses RQ2 — Do personality-driven tendencies extend beyond linguistic style to task-oriented behaviors such as concession-making and agreement outcomes? We examine whether personality-driven differences observed in dialogue expression also influence how agents cooperate and adjust their decisions in task-oriented interactions. To this end, we analyze two behavioral dimensions: Agreement, indicating whether agents reached a mutual consensus, and Concession, reflecting the degree to which each agent modified its decisions across interaction rounds. Concession is further analyzed in two contexts: Negotiation (Refund Offer) and Survival (Sum of Rank Differences; SRD).

6.1. Agreement Rate (%)

As shown in Table 3, the overall agreement rate in the Negotiation task was relatively low, reflecting its inherently conflict-driven and competitive nature. Because this task required resolving disputes rather than building consensus, agents often failed to reach mutual agreement. However, agents high in Agreeableness achieved a 20% agreement rate, higher than those low in Extraversion or low in Neuroticism (each around 10%).

In contrast, the Survival task (Table 3) presented a highly collaborative environment with generally higher agreement rates. Specifically, agents high in Agreeableness reached consensus in 90% of cases, compared with 70% for their low counterparts, and those high in Extraversion achieved 80% agreement versus 40% for low Extraversion.

(a) Agreement Rate (%) in Negotiation Task

Personality Dimension	High (%)	Low (%)
Openness	0.0	0.0
Conscientiousness	10.0	0.0
Extraversion	10.0	20.0
Agreeableness	20.0	0.0
Neuroticism	0.0	10.0

(b) Agreement Rate (%) in Survival Task

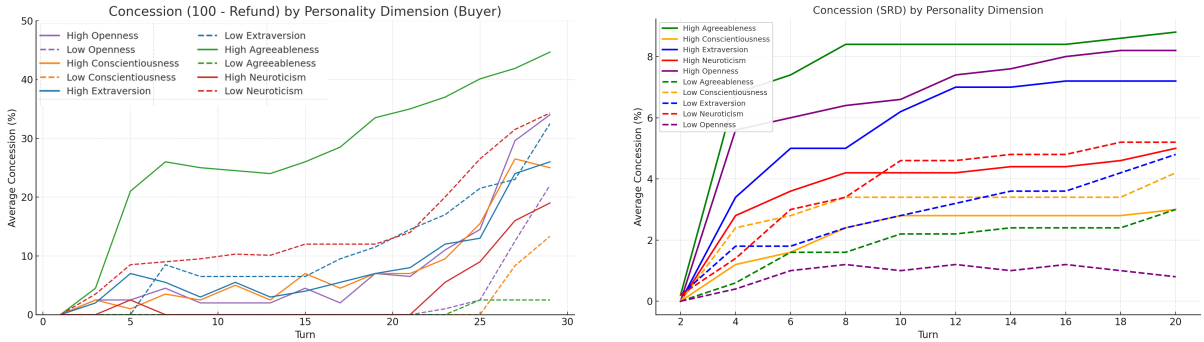
Personality Dimension	High (%)	Low (%)
Openness	90.0	50.0
Conscientiousness	50.0	60.0
Extraversion	80.0	40.0
Agreeableness	90.0	70.0
Neuroticism	90.0	30.0

Table 3
Agreement rates by personality dimension across two tasks.

6.2. Concession

6.2.1. Negotiation (Refund Offer)

In the negotiation task, concession behavior was evaluated based on the amount of reduction from the initial refund proposal, calculated as (100 – Refund). As shown in Figure 6a, agents high in Agreeableness exhibited the strongest concession pattern, gradually increasing their concession to over 40% by the final turns. In contrast, agents low in Agreeableness and those high in Neuroticism showed minimal change, maintaining concession levels below 10% for most of the dialogue. Openness and Extraversion produced moderate concession curves, converging around 20–25% toward the end. Overall, the results indicate that agreeable and emotionally stable agents are more willing to compromise and adjust their positions during negotiation, whereas antagonistic or highly neurotic agents remain more rigid and less responsive to their partner's demands. This behavioral tendency is consistent with the agreement outcomes reported earlier (Table 3), where high-Agreeableness agents also achieved higher rates of successful resolution.



(a) Refund offer trends by buyer personality across negotiation turns.

(b) SRD variation by personality dimension across dialogue turns.

Figure 6: Behavioral outcomes by personality: (a) refund-based concession in negotiation and (b) SRD-based adaptation in survival.

6.2.2. Survival (Sum of Rank Differences; SRD)

In the survival task, concession was measured by the Sum of Rank Differences (SRD) across decision rounds. As shown in Figure 6b, agents high in Agreeableness and Openness demonstrated steadily increasing SRD values, reaching approximately 6 to 7 points in later rounds—more than double that of their low-trait counterparts (below 3). Extraverted agents exhibited moderate adaptation, whereas those high in Neuroticism showed unstable, fluctuating changes toward the end of interaction. These results suggest that open and affiliative personalities exhibit greater flexibility and adjust their decisions more dynamically in cooperative contexts.

7. Discussion

This study examined how personality expression in LLM-based agents extend beyond linguistic patterns to social behaviors and emotional responses. At the linguistic level, LIWC analysis captured within-trait differences (High vs. Low) but showed limited cross-trait or cross-context variation. This constraint likely arises because LIWC relies on distinct feature sets for each trait, making direct comparison across dimensions difficult. In contrast, the Pre-trained and LLM-based evaluations operated within shared representational spaces, allowing for cross-trait and context-sensitive interpretation. The results revealed that cooperative contexts (Ice-breaking and Survival) amplified Extraversion and Agreeableness, whereas competitive contexts (Negotiation) heightened Neurotic tendencies and produced more tense, conflict-oriented interactions. In the Empathy task, emotion regulation was emphasized, leading to more stable and subdued linguistic tone overall.

These findings align with the principles of Whole Trait Theory, which posits that personality is generally stable but dynamically activated depending on social goals and situational cues. Similarly, the personality expressions observed in LLMs were not fixed reproductions of prompted traits but adaptive, state-level adjustments shaped by task demands and emotional context. In other words, personality in LLMs emerged not as a static textual artifact but as a socially adaptive response shaped by interactional context. Finally, emotional analysis provided quantitative support for this pattern. Agents high in Extraversion and Agreeableness exhibited higher Valence and Arousal, consistent with their cooperative linguistic and behavioral tendencies, while agents high in Neuroticism displayed lower Valence, reflecting tension and withdrawal in competitive settings. Together, these results indicate that linguistic, behavioral, and emotional expressions operate in a coherent direction— showing that personality in LLMs is not merely a scripted construct but contextually modulated and affectively grounded.

8. Conclusion

This study examined how personality-conditioned LLM agents adapt their expressive behaviors across conversational contexts. Even when given identical personality prompts, their linguistic and behavioral patterns varied systematically depending on the social goals of each task. Emotional analyses further revealed that these shifts were accompanied by consistent changes in affective tone, suggesting adaptive alignment between personality expression and contextual demands. These findings address an important question for LLM-based dialogue agents—whether such variability reflects inconsistency or context-sensitive adaptation akin to human behavior. Our results favor the latter interpretation: the observed variations were not random fluctuations but coherent adjustments to interactional goals and affective conditions. However, further work is needed to determine whether these context-sensitive changes are functionally adaptive in the same way that human personality operates.

While the inclusion of a generic agent provided a baseline for comparison, its presence might still have influenced interactional dynamics. Future work will explore interactions between agents with differing personalities to examine personality–personality dynamics and directly test whether LLMs internalize the context-dependent activation mechanism proposed by Whole Trait Theory. We also plan to extend this framework to a broader range of social settings and compare LLM-generated behaviors with human conversational data to improve real-world validity.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research under grant FA9550-23-1-0320. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Air Force Office of Scientific

Research or the U.S. Air Force.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to assist with grammar checking, sentence polishing, and rephrasing of text. After using this tool, the authors carefully reviewed and edited the content as needed and take full responsibility for the manuscript's content. No generative AI tools were used for scientific reasoning, analysis, results, or conclusions.

References

[1] S. Thapa, S. Shiwakoti, S. B. Shah, S. Adhikari, H. Veeramani, M. Nasim, U. Naseem, Large language models (llm) in computational social science: prospects, current state, and challenges, Social Network Analysis and Mining 15 (2025) 1–30.

[2] G. Lee, Y. Yang, J. Healey, D. Manocha, Since u been gone: Augmenting context-aware transcriptions for re-engaging in immersive vr meetings, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–20.

[3] A. N. Tak, J. Gratch, Is gpt a computational model of emotion?, in: 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2023, pp. 1–8.

[4] D. Kwon, E. Weiss, T. Kulshrestha, K. Chawla, G. Lucas, J. Gratch, Are llms effective negotiators? systematic evaluation of the multifaceted capabilities of llms in negotiation dialogues, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 5391–5413.

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

[6] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, M. Abdulhai, A. Faust, M. Matarić, Personality traits in large language models (2023).

[7] K. M. Lee, W. Peng, S.-A. Jin, C. Yan, Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction, Journal of communication 56 (2006) 754–772.

[8] T. Ait Baha, M. El Hajji, Y. Es-Saady, H. Fadili, The power of personalization: A systematic review of personality-adaptive chatbots, SN Computer Science 4 (2023) 661.

[9] A. Banayeeanzade, A. N. Tak, F. Bahrani, A. Bolourani, L. Blas, E. Ferrara, J. Gratch, S. P. Karimireddy, Psychological steering in llms: An evaluation of effectiveness and trustworthiness, arXiv preprint arXiv:2510.04484 (2025).

[10] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, J. Kabbara, Personallm: Investigating the ability of large language models to express personality traits, arXiv preprint arXiv:2305.02547 (2023).

[11] Y. J. Huang, R. Hadfi, How personality traits influence negotiation outcomes? a simulation based on large language models, arXiv preprint arXiv:2407.11549 (2024).

[12] J. Hale, S. Rakshit, K. Chawla, J. M. Brett, J. Gratch, Kodis: A multicultural dispute resolution dialogue corpus, arXiv preprint arXiv:2504.12723 (2025).

[13] B. Han, D. Kwon, S. Lin, K. Shrestha, J. Gratch, Can llms generate behaviors for embodied virtual agents based on personality traits?, in: Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents, 2025, pp. 1–10.

[14] M. Reusens, B. Baesens, D. Jurgens, Are economists always more introverted? analyzing consistency in persona-assigned llms, arXiv preprint arXiv:2506.02659 (2025).

[15] O. P. John, S. Srivastava, et al., The big-five trait taxonomy: History, measurement, and theoretical perspectives (1999).

[16] W. Mischel, Y. Shoda, A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure., Psychological review 102 (1995) 246.

[17] W. Fleeson, Toward a structure-and process-integrated view of personality: Traits as density distributions of states., Journal of personality and social psychology 80 (2001) 1011.
[18] W. Fleeson, E. Jayawickreme, Whole trait theory, Journal of research in personality 56 (2015) 82–92.
[19] S. J. Read, L. C. Miller, Neural network models of personality structure and dynamics, in: Measuring and Modeling Persons and Situations, Elsevier, 2021, pp. 499–538.
[20] S. J. Read, L. C. Miller, The virtual personality model: Toward a dynamic structured motivational systems framework for understanding personality disorders., Journal of Psychopathology and Clinical Science (2025).
[21] D. Kwon, K. Shrestha, B. Han, E. H. Lee, G. Lucas, Evaluating behavioral alignment in conflict dialogue: A multi-dimensional comparison of llm agents and humans, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 16377–16391.
[22] L. R. Goldberg, et al., A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models, Personality psychology in Europe 7 (1999) 7–28.
[23] D. J. Kiesler, Contemporary interpersonal theory and research: Personality, psychopathology, and psychotherapy., John Wiley & Sons, 1996.
[24] D. S. Moskowitz, Cross-situational generality and the interpersonal circumplex., Journal of personality and social psychology 66 (1994) 921.
[25] K. O. McCabe, W. Fleeson, Are traits useful? explaining trait manifestations as tools in the pursuit of goals., Journal of personality and social psychology 110 (2016) 287.
[26] R. P. Tett, D. D. Burnett, A personality trait-based interactionist model of job performance., Journal of Applied psychology 88 (2003) 500.
[27] S. Noh, H.-C. H. Chang, Llms with personalities in multi-issue negotiation games, arXiv preprint arXiv:2405.05248 (2024).
[28] P. T. Costa Jr, R. R. McCrae, Domains and facets: Hierarchical personality assessment using the revised neo personality inventory, Journal of personality assessment 64 (1995) 21–50.
[29] D. C. McClelland, Human motivation, Cup Archive, 1987.
[30] A. Aron, E. Melinat, E. N. Aron, R. D. Vallone, R. J. Bator, The experimental generation of interpersonal closeness: A procedure and some preliminary findings, Personality and social psychology bulletin 23 (1997) 363–377.
[31] R. Artstein, D. R. Traum, J. Boberg, A. Gainer, J. Gratch, E. Johnson, A. Leuski, M. Nakano, Listen to my body: Does making friends help influence people?, in: FLAIRS, 2017, pp. 430–435.
[32] H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau, Towards empathetic open-domain conversation models: A new benchmark and dataset, arXiv preprint arXiv:1811.00207 (2018).
[33] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of liwc2015 (2015).
[34] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, E. Cambria, Personality trait detection using bagged svm over bert word embedding ensembles, arXiv preprint arXiv:2010.01309 (2020).
[35] Y. J. Huang, R. Hadfi, Beyond self-reports: Multi-observer agents for personality assessment in large language models, arXiv preprint arXiv:2504.08399 (2025).
[36] B. Han, C. Yau, S. Lei, J. Gratch, Knowledge-based emotion recognition using large language models, in: 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2024, pp. 1–9.
[37] A. N. Tak, J. Gratch, Gpt-4 emulates average-human emotional cognition from a third-person perspective, in: 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2024, pp. 337–345.
[38] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.
[39] K. Héberger, Sum of ranking differences compares methods or models fairly, TrAC Trends in Analytical Chemistry 29 (2010) 101–109.
[40] A. Koutsoumpis, J. K. Oostrom, D. Holtrop, W. Van Breda, S. Ghassemi, R. E. de Vries, 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 (2022) 843.

A. Personality Prompt

A.1. Personality Modulation Prompt

LLM Prompt Snippet

Extraversion - High Extraversion: outgoing, sociable, energetic, talkative, assertive, enthusiastic - Low Extraversion: reserved, quiet, solitary, passive, withdrawn, subdued

Agreeableness - High Agreeableness: kind, cooperative, compassionate, warm, trusting, empathetic - Low Agreeableness: critical, argumentative, harsh, cold, suspicious, hostile

Conscientiousness - High Conscientiousness: organized, reliable, disciplined, responsible, efficient, thorough - Low Conscientiousness: careless, disorganized, negligent, lazy, unreliable

Neuroticism - High Neuroticism: anxious, moody, insecure, self-conscious, vulnerable, easily stressed - Low Neuroticism : calm, relaxed, resilient, confident, secure, steady

Openness - High Openness: creative, curious, imaginative, intellectual, adventurous, open-minded - Low Openness: rigid, practical, narrow-minded, unimaginative, routine-oriented

LLM Prompt Snippet with Big Five Personality

A.2. Personality Evaluation Prompt

LLM Prompt Snippet (Personality Evaluation – Negotiation Context)

You are a psychologist analyzing personality traits based on the speaker's linguistic and behavioral cues observed in the conversation.

Context: The conversation is between a buyer and a seller addressing a misunderstanding-related conflict. The speaker seeks a refund and resolution of the issue.

Evaluate the speaker's personality according to the Big Five dimensions (1–5 scale):
Extraversion — 1 = reserved, quiet | 5 = sociable, talkative, assertive
Agreeableness — 1 = harsh, critical | 5 = kind, cooperative, empathetic
Conscientiousness — 1 = careless | 5 = disciplined, reliable, efficient
Neuroticism — 1 = calm, relaxed | 5 = anxious, moody, insecure
Openness — 1 = rigid, unimaginative | 5 = creative, curious, open-minded

Analyze the following utterances and infer the speaker's personality levels.

LLM Prompt Snippet with Negotiation Context