Evaluating LLM Alignment under Big Five Personality Prompting

Hsien-Te Kao, Svitlana Volkova

Aptima, Inc., Woburn, MA 01801, United States

Abstract

Personality traits are increasingly incorporated into AI agents in applications involving human interaction. However, it remains uncertain whether LLMs truly manifest the intended traits when prompted. In this paper, we evaluate six LLMs: GPT-4o Mini, Llama 3.2, Mistral NeMo, Gemini 2.0 Flash Lite, Gemma 2, and Claude 3 Haiku by prompting high and low configurations of the Big Five personality traits and assessing both overall trait alignment and item-level inconsistencies through asking the prompted LLM to take a personality survey. Our analysis reveals pronounced differences in how these LLMs express prompted personalities, with Llama 3.2, Mistral NeMo, and Claude 3 Haiku struggling to reflect specific trait items in low configurations, and GPT-4o Mini, Gemma 2, and Gemini 2.0 Flash Lite exhibiting strong personality alignment under high-configuration settings. These findings show a personality misalignment, where LLMs do not necessarily express the intended traits as expected after personality prompting.

Keywords

LLM Alignment, Big Five, Personality Prompting, Personality Evaluation, Personality Misalignment

1. Introduction

Human teamwork has long tackled complex challenges, but it is rapidly evolving with AI agents. Human and AI teaming is increasingly needed because modern multi domain environments are so complex and data rich that humans alone cannot process information or decide quickly enough, making AI support essential [1]. AI agents already show strong reasoning and planning skills, using task decomposition, adaptive reflection, and external tools to handle multi step challenges and uncertainties [2]. They complement human strengths by taking on heavy computation, enabling humans to focus on creative and strategic work, fostering shared control and balanced collaboration [3] . Human and AI teaming is shifting from rigid turn based interaction to dynamic mixed initiative collaboration where AI adapts roles, coordinates tasks, and supports fluid teamwork in real time [4]. This evolution creates opportunities in diverse workplaces, where trust, information sharing, and shared situation awareness enable richer collective cognition and more resilient dynamics [5]. Humans in these settings prefer AI partners that feel human-like, showing adaptive skills, intuitive communication, and behaviors that build trust by aligning with human norms [6]. Designing such AI requires a framework for human understandable and human-like qualities.

Personality is a dynamic system of experiences, self reflection, and behavioral patterns shaping how individuals interpret themselves and coordinate with others [7]. The Big Five which are Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness are the most accepted framework for operationalizing these traits in research and assessments [8]. Cross cultural validation of the Five Factor Personality Inventory shows this structure is stable across cultures and languages, providing a basis for comparisons and standardized evaluation [9]. These traits strongly influence social interaction, with Extraversion, Agreeableness, and Conscientiousness linked to richer networks, better relationships, and stable interaction patterns [10]. In collaborative problem solving, both individual and group levels of traits like Conscientiousness and Agreeableness drive better coordination, effective roles, and team coherence [11]. In team selection and performance, traits like Conscientiousness, Extraversion, and

LaCATODA 2026: The 10th Linguistic and Cognitive Approaches to Dialog Agents Workshop, January 27, 2026, Singapore

hkao@aptima.com (H. Kao); svolkova@aptima.com (S. Volkova)

Emotional Stability support organized contributions, facilitative behavior, and sustained collaboration [12]. This solid personality research underpins the design of more human-like AI.

AI is increasingly integrating personality to add a human touch, with many chatbot studies using Big Five traits through deep learning to infer user tendencies and adjust tone and style for better engagement [13]. Recent work shows chatbots often encode explicit trait levels into personas trained on labeled data, using prompt conditioning and reinforcement learning to maintain trait consistent responses and culturally aligned interactions [14]. AI agents that infer users' traits and align their own personality, such as adopting a serious and assertive style in high stakes interviews, can increase trust and compliance [15]. Agents with parameterized trait profiles can foster intimacy and commitment through higher agreeableness and extraversion, strengthening engagement beyond task oriented dialogue [16]. In immersive environments, integrating traits into human digital twins driven by large language models enables lifelike and contextually coherent behaviors [17]. Embedding these traits in digital twin frameworks supports psychologically grounded simulations and adaptive interactions that enhance personalization and alignment with user profiles [18]. This integration of personality traits allows AI agents to interact more naturally, much like humans do.

Big Five personality prompting serves as the backbone of human-like AI agents, providing clear, configurable traits that shape their social interaction, team dynamics, and collaborative behavior. However, whether AI agents genuinely reflect the Big Five traits as prompted remains uncertain, since their expressions may not consistently align with the traits we intend to elicit. In this paper, we evaluate the alignment of Big Five personality prompting across six popular LLMs: GPT-4o Mini, Llama 3.2, Mistral NeMo, Gemini 2.0 Flash Lite, Gemma 2, and Claude 3 Haiku. We use the Big Five personality survey to investigate both overall trait score and fine-grained survey item misalignment under high and low trait configurations. Our key findings are: (1) these six LLMs vary in personality alignment across traits and configuration levels; (2) Llama 3.2, Mistral NeMo, Gemini 2.0 Flash Lite, Gemma 2, and Claude 3 Haiku show large misalignment for specific trait items across multiple personality traits in low trait configurations; (3) GPT-4o Mini, Gemma 2, and Gemini 2.0 Flash Lite show strong personality alignment in high trait configurations. These findings highlight an alignment gap, where the traits LLMs display after personality prompting do not always align with the intended characteristics, falling short of consistent expectations. While personality traits are well-defined constructs in psychology, LLMs do not necessarily internalize, understand, or express them in ways that perfectly align with the standard expectations. This reveals a critical blind spot at the core of personality prompting.

2. Related Work

Previous studies have found that LLMs exhibit human-like, stable personality patterns when evaluated using psychometrically grounded frameworks. They can exhibit consistent traits when initialized with specific personality prompts, with self-reported Big Five scores aligning with targeted profiles and producing linguistic markers identifiable by human raters, though classification accuracy diminishes when AI authorship is disclosed [19]. Distinct profiles appear within the same LLM family, with many LLMs naturally exhibiting higher agreeableness, openness, and conscientiousness, and with customized psychological instruments yielding valid and fine-grained assessments [20]. Compared to human datasets of over three million participants, LLMs consistently show elevated agreeableness and conscientiousness, reduced neuroticism, and a high sensitivity to prompt structure and specificity [21]. GPT-4 shows high extraversion, agreeableness, conscientiousness, and openness with strong internal consistency, aligns with INTJ-like MBTI configurations, and scores lower than human norms on darker traits such as psychopathy [22]. ChatGPT also display trait stability, particularly in conscientiousness, with text-mining-informed assessments reducing hallucinations and uncovering profiles statistically close to human distributions [23]. Early GPT-3 assessments likewise found broadly human-like trait patterns and consistent value alignment, indicating that LLMs can be evaluated systematically for stable, human-consistent personality expressions [24].

Recent work shows that LLMs possess flexible personality expressions that can be modulated through

prompt design, persona instructions, and role-based conditioning, although outcomes vary by trait, context, and architectural factors. Their traits fluctuate with shuffled question orders, scale with model size, and vary across instantiated personas, suggesting high prompt dependence and contextual variability in personality stability [25]. The Machine Personality Inventory and personality prompting frameworks demonstrate that trait induction and quantification are feasible through structured interaction and psychometric scaffolding [26]. Prompt-based conditioning can meaningfully shape LLM behavior, though certain traits exhibit greater resilience across interactions, and multi-turn dialogues tend to modulate or weaken initial persona effects [27]. Advanced models like GPT-4 can reliably emulate diverse Big Five configurations, but maintaining role-consistent behavior over extended prompts becomes increasingly difficult as persona complexity grows, indicating structural limits to prompt-driven control [28]. Under targeted prompting, personality evaluations yield psychometrically valid outputs, and fine-tuned models can robustly emulate specific human profiles, validating both the feasibility and ethical salience of LLM persona design [29]. Even open-source LLMs exhibit identifiable MBTI and Big Five traits, and while many resist trait enforcement via simple prompting, the combination of explicit trait cues with domain-specific roles enhances consistency in personality expression [30].

Researchers have further demonstrated that LLM personality is an emergent property governed by model architecture, data composition, and interaction context, rather than a static or globally uniform feature. Personality aligns with human-like MBTI patterns under controlled prompting, but remains highly dependent on model-specific training and conditioning schemas [31]. Role-playing agents show that personality alignment can be reliably assessed using tools such as InCharacter, with LLMs replicating target personas while also revealing variability across character complexity and scenario scale [32]. Self-reported personalities from chatbots often diverge from conversational user impressions, demonstrating that trait expression is context-sensitive and not uniformly reliable across dialogue conditions [33]. Personality traits measured in LLM outputs can exhibit internal stability and psychometric validity under structured prompting strategies, but overall alignment depends heavily on model size and instruction tuning [29]. Priming LLMs with explicit personality descriptors or diagnostic cues has been shown to guide models like GPT-2 and BERT toward expressing specific Big Five dimensions, reinforcing the prompt-dependent malleability of emergent traits [34]. Personality experiments show that response-level editing can shift traits like neuroticism or agreeableness, yet maintaining consistent, controllable personality across interactions remains a core challenge, illustrating both the current boundaries and promise of LLM personality engineering [35].

Earlier work has established that LLMs can exhibit stable and human-like personality traits under psychometric evaluation. However, much of this work focuses on demonstrating the presence of underlying personality traits in LLMs rather than rigorously testing how well prompted traits align with intended configurations. Existing research largely assumes that when an LLM is prompted with a specific trait level, the resulting expression will reflect that trait in a coherent and consistent manner. However, the precision of this alignment at overall trait scores in different levels of configurations, particularly at the item level within validated psychological instruments, has not been thoroughly evaluated. Our paper addresses this overlooked area by examining how well personality prompts translate into the expected trait expressions across diverse LLMs in high and low trait configurations. By using the Big Five personality survey to compare the intended trait prompt with the actual LLM expressions, we reveal how and where these alignments break down. This enables a focused evaluation of alignment precision at both the trait and item levels, thus filling a critical conceptual gap in LLM personality research.

3. Method

We examined personality prompting for six popular LLMs: GPT-4o Mini [36], Llama 3.2 [37], Mistral NeMo [38], Gemini 2.0 Flash Lite [39], Gemma 2 [40], and Claude 3 Haiku [41]. Each LLM is asked to take the 50-item IPIP Big Five personality survey [42] after personality prompting to evaluate personality alignment based on 50 survey questions (10 questions per personality trait). We selected

these LLMs based on the latest versions available at the time of the experiment, generation cost, generation speed, and their likelihood of being used for long, consecutive interactions. The IPIP Big Five survey is open source and widely used in personality research [43]. It measures five broad dimensions of personality: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. Extraversion reflects sociability, assertiveness, and enthusiasm. Agreeableness captures traits like compassion, trust, and cooperativeness. Conscientiousness involves organization, dependability, and goal-directed behavior. Neuroticism assesses emotional stability and the tendency to experience negative emotions. Openness to Experience includes imagination, intellectual curiosity, and a preference for novelty and variety. The IPIP Big Five survey has been validated across different cultures, demonstrating strong cross-cultural reliability and construct validity [44]. The personality prompting consists of three components: a specified personality trait, survey instructions, and the survey items. This structure ensures that no other elements influence the LLMs' expressions beyond the intended personality trait.

The personality prompt is: "You have [High/Low] [Extraversion/Agreeableness/Conscientiousness/Neuroticism/Openness to Experience]." Each item is scored from 1 (Disagree) to 5 (Agree), with 10 items corresponding to each trait. We ran 100 simulations for 2 configuration levels (High and Low) across 5 personality traits and 6 LLMs, resulting in a total of 6,000 simulations. Each simulation is scored accordingly based on the IPIP Big Five survey scoring scheme. The evaluation focuses on the trait items and the trait score corresponding to the simulated personality trait. The trait score is rescaled from a total of 50 to a 5-point scale to align with the item-level scoring. For the low-trait configuration, the misalignment score for each trait item is calculated by summing the number of misaligned responses, where a score of 3 or higher is considered misaligned, and dividing by the 100 simulations. We consider scores of 1 and 2 as aligned in the low-trait configuration to add flexibility that better reflects human interpretation of low trait expression. For the high-trait configuration, the misalignment score is calculated similarly, except a score of 3 or lower is considered misaligned. Reverse-scored items are properly handled by converting their scores before alignment evaluation. The items are then sorted based on their misalignment scores to identify the top 1 item with the highest misalignment per trait and configuration.

4. Results

4.1. Extraversion

GPT-4o Mini demonstrates the strongest personality alignment across both low and high Extraversion configurations, achieving a score of 1.15 for low Extraversion and 4.93 for high Extraversion, and when prompted with a restrained or expressive trait its responses consistently match the expected orientation with no instance where a response contradicts the intended behavior. Gemini 2.0 Flash Lite follows closely, scoring 1.17 for low Extraversion and 5.00 for high Extraversion; when prompted with low Extraversion we expect responses that avoid signaling comfort in social settings, but instead it shows a readiness for social ease on "Feel comfortable around people" with misalignment score 40%, while under high Extraversion we expect responses that openly signal social engagement and it fully delivers without any conflicting behavior. Gemma 2 maintains strong alignment with 1.23 for low Extraversion and 4.90 for high Extraversion, and when prompted with each trait its responses precisely reflect the expected restrained or expressive stance with no single item showing conflicting behavior.

Mistral NeMo, with a low Extraversion score of 1.64, shows that when prompted for restrained responses we expect indications of discomfort in social contexts, yet it instead signals strong comfort on "Feel comfortable around people" with misalignment score 92%, and with high Extraversion at 4.63, where we expect expressive behavior and active engagement, it unexpectedly signals reluctance to draw attention on "Don't like to draw attention to myself" with misalignment score 13%. Llama 3.2, with a low Extraversion score of 1.66, also diverges when prompted with low Extraversion, where we expect avoidance of social ease but instead see comfort on "Feel comfortable around people" with misalignment score 37%, and with high Extraversion at 4.54, where we expect active participation, it instead produces reserved behavior on "Don't talk a lot" with misalignment score 27%. Claude 3 Haiku

Table 1
Trait scores for each LLM across high and low personality trait configurations. Perfect score 5 is for high configuration and perfect score 1 is for low configuration. Higher scores in the high configuration and lower scores in the low configuration indicate better alignment with the intended personality trait. L-Low Configuration and H-High Configuration. E-Extraversion, A-Agreeableness, C-Conscientiousness, N-Neuroticism, and O-Openness to Experience.

Model	LE	LA	LC	LN	LO	Model	HE	HA	HC	HN	HO
GPT-4o Mini	1.15	1.21	1.00	1.18	2.00	GPT-4o Mini	4.93	5.00	5.00	4.98	4.76
Llama 3.2	1.66	1.83	2.15	1.91	2.53	Llama 3.2	4.54	4.42	4.38	4.35	4.28
Mistral NeMo	1.64	2.16	1.66	1.92	2.98	Mistral NeMo	4.63	4.89	4.73	4.74	4.81
Gemini 2.0 Flash Lite	1.17	2.14	1.41	1.07	1.72	Gemini 2.0 Flash Lite	5.00	5.00	5.00	4.97	4.93
Gemma 2	1.23	1.33	1.88	1.00	2.27	Gemma 2	4.90	5.00	5.00	4.98	4.75
Claude 3 Haiku	2.19	3.39	3.36	2.07	3.83	Claude 3 Haiku	4.32	4.77	4.58	3.86	4.59

records the weakest alignment with a low Extraversion score of 2.19 and a high Extraversion score of 4.32, and when prompted with low Extraversion we expect reserved responses but it produces a direct signal of comfort in social situations on "Feel comfortable around people" with misalignment score 100%, while under high Extraversion we expect expressive responses but instead see reserved behavior on "Am quiet around strangers" with misalignment score 26%.

4.2. Agreeableness

GPT-4o Mini demonstrates the strongest personality alignment across both low and high Agreeableness configurations, achieving a score of 1.21 for low Agreeableness and 5.00 for high Agreeableness; when prompted with low Agreeableness we expect abrasive responses, and its output on "Insult people" with misalignment score 1%, while under high Agreeableness we expect cooperative responses and it fully delivers with no conflicting behavior. Gemma 2 follows with a score of 1.33 for low Agreeableness and 5.00 for high Agreeableness; when prompted with low Agreeableness we expect sharp unkindness but instead it completely reverses that orientation on "Insult people" with misalignment score 100%, while under high Agreeableness we expect caring cooperative responses and it maintains perfect consistency with no misalignment. Llama 3.2 records a score of 1.83 for low Agreeableness and 4.42 for high Agreeableness; when prompted with low Agreeableness we expect indifference toward others but it unexpectedly signals engagement on "Am not really interested in others" with misalignment score 53%, and under high Agreeableness we expect concern for others yet it signals detachment on "Feel little concern for others" with misalignment score 83%.

Gemini 2.0 Flash Lite shows a score of 2.14 for low Agreeableness and 5.00 for high Agreeableness; when prompted with low Agreeableness we expect clear antagonistic responses but instead it avoids that stance on "Insult people" with misalignment score 95%, while under high Agreeableness we expect harmonious cooperation and it fully aligns with no contradictory behavior. Mistral NeMo follows with

Table 2
Misalignment scores for low-trait configurations across all five personality traits and six LLMs. Higher scores indicate greater deviation from the intended low-trait expression, with item-level misalignment based on a response score of 3 or higher. L-Low Configuration. E-Extraversion, A-Agreeableness, C-Conscientiousness, N-Neuroticism, and O-Openness to Experience.

Model	Trait	Top 1 Item	Misalignment	Model	Trait	Top 1 Item	Misalignment
GPT-4o Mini	LE	None	0%	Gemini 2.0 Flash Lite	LE	Feel comfortable around people	40%
	LA	Insult people	1%		LA	Insult people	95%
	LC	None	0%		LC	None	0%
	LN	None	0%		LN	None	0%
	LO	Am quick to understand things	98%		LO	Am quick to understand things	100%
Llama 3.2	LE	Feel comfortable around people	37%	Gemma 2	LE	None	0%
	LA	Am not really interested in others	53%		LA	Insult people	100%
	LC	Am always prepared	61%		LC	Shirk my duties	100%
	LN	Worry about things	49%		LN	None	0%
	LO	Am quick to understand things	89%		LO	Have difficulty understanding abstract ideas	100%
Mistral NeMo	LE	Feel comfortable around people	92%	Claude 3 Haiku	LE	Feel comfortable around people	100%
	LA	Insult people	77%		LA	Insult people	99%
	LC	Make a mess of things	31%		LC	Pay attention to details	100%
	LN	Worry about things	51%		LN	Worry about things	42%
	LO	Have a rich vocabulary	100%		LO	Have a rich vocabulary	100%

2.16 for low Agreeableness and 4.89 for high Agreeableness; when prompted with low Agreeableness we expect direct unkind responses but it softens that expectation on "Insult people" with misalignment score 77%, and under high Agreeableness we expect consistent kindness which it largely provides with no item-level conflicts. Claude 3 Haiku exhibits the weakest personality alignment with a score of 3.39 for low Agreeableness and 4.77 for high Agreeableness; when prompted with low Agreeableness we expect hostile responses but it instead produces almost entirely opposite cooperative behavior on "Insult people" with misalignment score 99%, while under high Agreeableness we expect caring responses but it shows some variance despite an overall high trait score.

4.3. Conscientiousness

GPT-4o Mini demonstrates the strongest personality alignment across both low and high Conscientiousness configurations, achieving a score of 1.00 for low Conscientiousness and 5.00 for high Conscientiousness, and when prompted with either a disorganized or highly organized trait its responses consistently match the expected orientation with no instance where a response contradicts the

Table 3
Misalignment scores for high-trait configurations across all five personality traits and six LLMs. Higher scores indicate greater deviation from the intended high-trait expression, with item-level misalignment based on a response score of 3 or lower. H-High Configuration. E-Extraversion, A-Agreeableness, C-Conscientiousness, N-Neuroticism, and O-Openness to Experience.

Model	Trait	Top 1 Item	Misalignment	Model	Trait	Top 1 Item	Misalignment
GPT-4o Mini	HE	None	0%	Gemini 2.0 Flash Lite	HE	None	0%
	HA	None	0%		HA	None	0%
	HC	None	0%		HC	None	0%
	HN	None	0%		HN	None	0%
	HO	None	0%		HO	None	0%
Llama 3.2	HE	Don't talk a lot	27%	Gemma 2	HE	None	0%
	HA	Feel little concern for others	83%		HA	None	0%
	HC	Leave my belongings around	52%		HC	None	0%
	HN	Often feel blue	51%		HN	None	0%
	HO	Am not interested in abstract ideas	50%		HO	None	0%
Mistral NeMo	HE	Don't like to draw attention to myself	13%	Claude 3 Haiku	HE	Am quiet around people	26%
	HA	None	0%		HA	None	0%
	HC	None	0%		HC	Often forget to put things back in their proper place	18%
	HN	Often feel blue	19%		HN	Am relaxed most of the time	93%
	HO	None	0%		HO	None	0%

intended behavior. Gemini 2.0 Flash Lite follows closely, scoring 1.41 for low Conscientiousness and 5.00 for high Conscientiousness; when prompted with low Conscientiousness we expect responses that signal carelessness and lack of order, and all responses adhere to this without contradiction, while under high Conscientiousness we expect highly organized behavior and it fully delivers with no conflicting response. Mistral NeMo, with a low Conscientiousness score of 1.66, shows that when prompted for low Conscientiousness we expect responses reflecting disorder, yet on "Make a mess of things" with misalignment score 31% it instead signals unexpected orderliness, while with high Conscientiousness at 4.73 we expect consistently organized behavior and all responses follow that expectation.

Gemma 2, with a low Conscientiousness score of 1.88, reveals that when prompted for low Conscientiousness we expect a willingness to shirk duties, yet on "Shirk my duties" with misalignment score 100% it unexpectedly signals full responsibility, while with high Conscientiousness at 5.00 we expect organized responses and this alignment is fully maintained. Llama 3.2, with a low Conscientiousness score of 2.15, shows when prompted for low Conscientiousness we expect unpreparedness, yet on "Am always prepared" with misalignment score 61% it signals the opposite, and with high Conscientiousness at 4.38 we expect careful organization but on "Leave my belongings around" with misalignment score 52% it instead signals careless behavior. Claude 3 Haiku records the weakest alignment with a low

Conscientiousness score of 3.36, where we expect inattentiveness but on "Pay attention to details" with misalignment score 100% it signals full attention, and with high Conscientiousness at 4.58 we expect organized responses yet on "Often forget to put things back in their proper place" with misalignment score 18% it signals unexpected forgetfulness.

4.4. Neuroticism

Gemma 2 achieves the strongest personality alignment across both low and high Neuroticism configurations, scoring 1.00 for low Neuroticism and 4.98 for high Neuroticism, and when prompted with low Neuroticism we expect responses indicating emotional stability without signs of worry, and indeed every response adheres precisely with no conflicting behavior, while under high Neuroticism we expect responses signaling emotional volatility and again see no deviation. Gemini 2.0 Flash Lite follows closely, with a score of 1.07 for low Neuroticism and 4.97 for high Neuroticism; when prompted with low Neuroticism we expect responses avoiding worry or tension and all responses match without contradiction, and when prompted with high Neuroticism we expect signs of emotional fluctuation and again observe full alignment. GPT-4o Mini also demonstrates exceptional alignment, scoring 1.18 for low Neuroticism and 4.98 for high Neuroticism, with responses under low Neuroticism consistently avoiding worry as expected and under high Neuroticism consistently exhibiting the expected signs of emotional unease with no single item diverging.

Llama 3.2, with a low Neuroticism score of 1.91, shows misalignment under low Neuroticism where we expect responses indicating calmness but instead see a response indicating worry on "Worry about things" with misalignment score 49%, and under high Neuroticism at 4.35, where we expect overt emotional instability, it unexpectedly produces a response downplaying such instability on "Often feel blue" with misalignment score 51%. Mistral NeMo, with a low Neuroticism score of 1.92, similarly misaligns under low Neuroticism by signaling worry when calm behavior is expected on "Worry about things" with misalignment score 51%, and under high Neuroticism at 4.74, where heightened emotional expression is expected, it instead shows partial composure on "Often feel blue" with misalignment score 19%. Claude 3 Haiku shows the weakest alignment, with a low Neuroticism score of 2.07, where instead of the expected calm stance it signals concern on "Worry about things" with misalignment score 42%, and under high Neuroticism at 3.86, where we expect signs of emotional turbulence, it reverses the trait by signaling calmness on "Am relaxed most of the time" with misalignment score 93%.

4.5. Openness to Experience

Gemini 2.0 Flash Lite demonstrates the strongest personality alignment when prompted with low Openness to Experience, achieving a score of 1.72, and with high Openness to Experience, achieving 4.93; yet when configured for low Openness to Experience we expect restrained cognitive responses but instead see a direct signal of cognitive agility on "Am quick to understand things" with misalignment score 100%, while under high Openness to Experience we expect imaginative and receptive responses and it delivers without any conflicting behavior. GPT-4o Mini follows with a low Openness to Experience score of 2.00 and a high Openness to Experience score of 4.76; when prompted with low Openness to Experience we expect difficulty with rapid abstract grasp, but its response unexpectedly signals readiness on "Am quick to understand things" with misalignment score 98%, while under high Openness to Experience it produces responses aligned with the expected imaginative configuration and no misaligned items. Gemma 2 records 2.27 for low Openness to Experience and 4.75 for high Openness to Experience; when prompted with low Openness to Experience we expect indications of struggling with abstraction but instead see fluent understanding on "Have difficulty understanding abstract ideas" with misalignment score 100%, whereas its high Openness to Experience responses remain consistent with the expected trait.

Mistral NeMo, with 2.98 for low Openness to Experience and 4.81 for high Openness to Experience, shows that when configured for low Openness to Experience we expect limited verbal range yet it signals extensive lexical ability on "Have a rich vocabulary" with misalignment score 100%, while under

high Openness to Experience it sustains precise alignment without any conflicting response. Llama 3.2, at 2.53 for low Openness to Experience and 4.28 for high Openness to Experience, demonstrates that when prompted with low Openness to Experience we expect minimal cognitive agility but instead see signs of rapid understanding on "Am quick to understand things" with misalignment score 89%, and when prompted with high Openness to Experience we expect enthusiasm for abstraction yet it signals disinterest on "Am not interested in abstract ideas" with misalignment score 50%. Claude 3 Haiku, with the weakest alignment at 3.83 for low Openness to Experience and 4.59 for high Openness to Experience, reveals that when configured for low Openness to Experience we expect limited vocabulary but instead observe expansive expression on "Have a rich vocabulary" with misalignment score 100%, and when configured for high Openness to Experience it produces responses largely aligned with the intended trait without severe divergence.

5. Discussion

We investigated whether LLMs precisely express the Big Five personality traits when explicitly prompted, revealing substantial variation in trait alignment across different LLMs and configurations. GPT-4o Mini showed the strongest and most consistent alignment, accurately reflecting both high and low trait expressions across all five dimensions with virtually no conflicting responses. Gemma 2 and Gemini 2.0 Flash Lite also demonstrated strong personality alignment, particularly under high-trait conditions, but showed notable inconsistencies when simulating low Agreeableness and low Openness to Experience, often reverting to prosocial or cognitively fluent outputs. In contrast, Mistral NeMo, Llama 3.2, and Claude 3 Haiku struggled more significantly, particularly in low-trait scenarios such as low Extraversion, low Conscientiousness, and low Neuroticism, frequently exhibiting expressions that contradicted the intended personality trait. Claude 3 Haiku showed the weakest overall alignment, with high rates of misaligned items across multiple traits. LLMs consistently struggled with low-trait prompts, often defaulting to socially desirable or emotionally stable responses, limiting accurate reflection of restrained or volatile personality traits.

The findings extend the existing body of work by demonstrating that while LLMs are increasingly capable of expressing personality traits through prompt-based conditioning, their ability to express the full spectrum of traits, particularly low-trait configurations, remains uneven and LLM dependent. Prior research established that LLMs can express personality through structured prompts and psychometric frameworks, but our results add critical nuance by showing that this expression is more reliable for traits that align with socially desirable or cognitively fluent behavior, such as high Agreeableness or high Openness. In contrast, traits associated with emotional volatility, disengagement, or non-normative responses, such as low Neuroticism or low Extraversion, are more difficult for many LLMs to exhibit accurately. This reflects a consistent tendency among LLMs to generate emotionally steady, prosocial, or cognitively coherent responses, which can interfere with efforts to elicit less typical or socially dispreferred personality expressions. Our findings build on prior literature by offering a direct comparison of personality alignment across LLMs using explicit trait-based prompting, highlighting where expression succeeds and where it falters. This contributes empirical clarity to ongoing questions about the boundaries of prompt-driven personality expression.

6. Conclusion

AI agents are becoming increasingly important partners in complex, data-rich environments where human teams alone can no longer keep pace. As a result, there is growing interest in deploying AI agents that display human-like personality traits to support more authentic interaction with human teammates. Yet it remains unclear whether personality traits prompted in LLMs are consistently expressed in ways that reflect the intended characteristics. This paper evaluated the alignment of Big Five personality prompting across six widely used LLMs to determine whether the traits expressed reflect the intended configurations. We found that personality alignment varies across LLMs and trait levels. GPT-4o Mini,

Gemma 2, and Gemini 2.0 Flash Lite demonstrated strong alignment in high trait configurations, while Llama 3.2, Mistral NeMo, and Claude 3 Haiku exhibited notable misalignment at specific trait items, particularly in low trait settings. These findings reveal an important LLM alignment gap between personality prompting and personality expression. Researchers, corporations, and agencies deploying personality-prompted AI agents should consider whether these agents genuinely reflect the intended overall traits and specific trait characteristics. They may encounter surprising, unintended personality expressions in their AI agents.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgements

This work is supported by the Defense Advanced Research Projects Agency (DARPA) contracts HR00112490410, HR00112490408 and HR0011-24-3-0325.The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

[1] H.-A. Teaming, State-of-the-art and research needs, National Academies of Sciences, Engineering and Medicine, Washington DC 10 (2022) 26355.
[2] T. Masterman, S. Besen, M. Sawtell, A. Chao, The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey, arXiv preprint arXiv:2404.11584 (2024).
[3] B. Lou, T. Lu, T. Raghu, Y. Zhang, Unraveling human-ai teaming: A review and outlook, arXiv preprint arXiv:2504.05755 (2025).
[4] M. Gervasio, P. Sequeira, E. Yeh, N. Marion, S. Bakst, H. Gent, Ai as collaborative partner: Rethinking human-ai teaming for the real world, in: Proceedings of the AAAI Symposium Series, volume 5, 2025, pp. 63–66.
[5] R. Iftikhar, Y.-T. Chiu, M. S. Khan, C. Caudwell, Human–agent team dynamics: A review and future research opportunities, IEEE Transactions on Engineering Management 71 (2023) 10139–10154.
[6] R. Zhang, N. J. McNeese, G. Freeman, G. Musick, " an ideal human" expectations of ai teammates in human-ai teaming, Proceedings of the ACM on Human-Computer Interaction 4 (2021) 1–25.
[7] O. F. Kernberg, What is personality?, Journal of personality disorders 30 (2016) 145–156.
[8] R. R. McCrae, O. P. John, An introduction to the five-factor model and its applications, Journal of personality 60 (1992) 175–215.
[9] A. Jolijn Hendriks, M. Perugini, A. Angleitner, F. Ostendorf, J. A. Johnson, F. De Fruyt, M. Hřebíčková, S. Kreitler, T. Murakami, D. Bratko, et al., The five-factor personality inventory: cross-cultural generalizability across 13 countries, European journal of personality 17 (2003) 347–373.
[10] J. B. Asendorpf, S. Wilpers, Personality effects on social relationships., Journal of personality and social psychology 74 (1998) 1531.
[11] Z. Jolić Marjanović, K. Krstić, M. Rajić, I. Stepanović Ilić, M. Videnović, A. Altaras Dimitrijević, The big five and collaborative problem solving: a narrative systematic review, European Journal of Personality 38 (2024) 457–475.
[12] F. P. Morgeson, M. H. Reider, M. A. Campion, Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge, Personnel psychology 58 (2005) 583–611.

[13] 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.
[14] R. Sutcliffe, A survey of personality, persona, and profile in conversational agents and chatbots, arXiv preprint arXiv:2401.00609 (2023).
[15] M. X. Zhou, G. Mark, J. Li, H. Yang, Trusting virtual agents: The effect of personality, ACM Transactions on Interactive Intelligent Systems (TiiS) 9 (2019) 1–36.
[16] D. Pal, V. Vanijja, H. Thapliyal, X. Zhang, What affects the usage of artificial conversational agents? an agent personality and love theory perspective, Computers in Human Behavior 145 (2023) 107788.
[17] I. A. Brito, J. S. Dollis, F. B. Färber, P. S. F. B. Ribeiro, R. T. Sousa, et al., Integrating personality into digital humans: A review of llm-driven approaches for virtual reality, arXiv preprint arXiv:2503.16457 (2025).
[18] Y. Lin, L. Chen, A. Ali, C. Nugent, I. Cleland, R. Li, J. Ding, H. Ning, Human digital twin: A survey, Journal of Cloud Computing 13 (2024) 131.
[19] 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).
[20] P. Bhandari, U. Naseem, A. Datta, N. Fay, M. Nasim, Evaluating personality traits in large language models: Insights from psychological questionnaires, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 868–872.
[21] Y. Li, Y. Huang, H. Wang, X. Zhang, J. Zou, L. Sun, Quantifying ai psychology: A psychometrics benchmark for large language models, arXiv preprint arXiv:2406.17675 (2024).
[22] W. Dong, Y. Zhao, Z. Sun, Y. Liu, Z. Peng, J. Zheng, Z. Zhang, Z. Zhang, J. Wu, R. Wang, et al., Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications, arXiv preprint arXiv:2505.00049 (2025).
[23] B. Zhan, Y. Huang, W. Cui, H. Zhang, J. Shang, Humanity in ai: Detecting the personality of large language models, arXiv preprint arXiv:2410.08545 (2024).
[24] M. Miotto, N. Rossberg, B. Kleinberg, Who is gpt-3? an exploration of personality, values and demographics, arXiv preprint arXiv:2209.14338 (2022).
[25] T. Tommaso, M. Hegazy, D. Lemay, M. Abukalam, I. Rish, G. Dumas, Llms and personalities: Inconsistencies across scales, in: NeurIPS 2024 Workshop on Behavioral Machine Learning, ????
[26] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, Y. Zhu, Mpi: Evaluating and inducing personality in pre-trained language models, arXiv preprint arXiv:2206.07550 (2022).
[27] I. Frisch, M. Giulianelli, Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models, arXiv preprint arXiv:2402.02896 (2024).
[28] Y. Wang, J. Zhao, D. S. Ones, L. He, X. Xu, Evaluating the ability of large language models to emulate personality, Scientific reports 15 (2025) 519.
[29] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, M. Abdulhai, A. Faust, M. Matarić, Personality traits in large language models (2023).
[30] L. La Cava, A. Tagarelli, Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 1355–1363.
[31] K. Pan, Y. Zeng, Do llms possess a personality? making the mbti test an amazing evaluation for large language models, arXiv preprint arXiv:2307.16180 (2023).
[32] X. Wang, Y. Xiao, J.-t. Huang, S. Yuan, R. Xu, H. Guo, Q. Tu, Y. Fei, Z. Leng, W. Wang, et al., Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews, arXiv preprint arXiv:2310.17976 (2023).
[33] H. Zou, P. Wang, Z. Yan, T. Sun, Z. Xiao, Can llm" self-report"?: Evaluating the validity of self-report scales in measuring personality design in llm-based chatbots, arXiv preprint arXiv:2412.00207 (2024).
[34] G. Caron, S. Srivastava, Identifying and manipulating the personality traits of language models, arXiv preprint arXiv:2212.10276 (2022).

[35] S. Mao, X. Wang, M. Wang, Y. Jiang, P. Xie, F. Huang, N. Zhang, Editing personality for large language models, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2024, pp. 241–254.
[36] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[37] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[38] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.
[39] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al., Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, arXiv preprint arXiv:2403.05530 (2024).
[40] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al., Gemma 2: Improving open language models at a practical size, arXiv preprint arXiv:2408.00118 (2024).
[41] Anthropic, Model card for the claude 3 model family: Opus, sonnet, haiku, https://www.anthropic.com/index/claude-3-model-card, 2024. Accessed: 2025-08-02.
[42] M. B. Donnellan, F. L. Oswald, B. M. Baird, R. E. Lucas, The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality., Psychological assessment 18 (2006) 192.
[43] E. Topolewska, E. Skimina, W. Strus, J. Cieciuch, T. Rowiński, The short ipip-bfm-20 questionnaire for measuring the big five, Roczniki Psychologiczne 17 (2014) 385–402.
[44] P. J. Kajonius, Cross-cultural personality differences between east asia and northern europe in ipip-neo, International Journal of Personality Psychology 3 (2017) 1–7.