scientific reports

Check for updates

OPEN

LLM-based robot personality simulation and cognitive system

Jia-Hsun Lo[1], Han-Pang Huang[1]✉ & Jie-Shih Lo[2]

The inherence of personality in human-robot interaction enhances conversational dynamics and user experience. The deployment of Chat GPT-4 within a cognitive robot framework is designed by using state-space realization to emulate specific personality traits, incorporating elements of emotion, motivation, visual attention, and both short-term and long-term memory. The encoding and retrieval of long-term memory are facilitated through document embedding techniques, while emotions are generated based on predictions of future events. This framework processes textual and visual information, responding or initiating actions in accordance with the configured personality settings and cognitive processes. The constancy and effectiveness of the personality simulation have been compared to human baseline and validated via two personality assessments: the International Personality Item Pool – Neuroticism, Extraversion and Openness (IPIP-NEO) and the Big Five personality test. Our proposed personality model of cognitive robot is designed by using Kelly's role construct repertory, Cattell's 16 personality factors and preferences, which are analyzed by construct validity and compared to human subjects. Theory of mind is observed in personality simulation, which perform better second-order of belief compared to other agent on the improved theory of mind dataset (ToMi dataset). Based on the proposed methods, our designed robot, Mobi, is enable to chat based on its own personality, handle social conflicts and understand user's intent. Such simulations can achieve a high degree of human likeness, characterized by conversations that are flexible and imbued with intention.

Keywords Human robot interaction, Large language model, Robot personality and cognition, Theory of mind

Personality has been identified as a crucial factor in understanding the quality of robot deployments in organizations and in broader society[1]. Although the uncanny valley hypothesis posits that humans feel uncomfortable in the presence of robots with human-like features, this unsettling first impression is significantly altered through interaction[2]. The personality of a robot is considered preferable and is associated with desirable social responses[3]. To determine a suitable personality for a robot, prior research suggests that preferences for robot personalities may indeed vary depending on the context of the robot's role and the stereotypical perceptions people hold for certain occupations[4]. It has been demonstrated that users are capable of distinguishing between robot personalities, which results in differing preferences between goal-oriented and experience-oriented scenarios[5].

A robot has been designed to exhibit traits of introversion and extroversion to assist post-stroke individuals in their rehabilitation exercises[6]. Research indicates that a socially assistive robot's autonomous behavior, when adapted to the user's personality, can enhance task performance. Studies have demonstrated that a robot's extroversion and dominance influence people's perceptions of its intelligence, social capabilities, and likability[7]. Furthermore, the robot's personality has been shown to improve human-robot interactions; users report greater enjoyment when interacting with a robot whose personality complements their own, as opposed to one with a similar personality[8]. Guidelines for the effective design of service robots have been proposed to elicit desired emotional responses from users[9]. The incorporation of personality into robots facilitates human-robot interactions with greater social presence, leading to outcomes such as increased acceptance and heightened emotional engagement during service encounters. Previous research has primarily employed the Big Five model to instill personality in robots[10]. In contrast, interpersonal theory emphasizes interactions with others rather than internal characteristics[11], diverging from the Big Five model. Moreover, the assessment of robot personality has been calibrated using the International Personality Item Pool (IPIP) and the Big Five Inventory (BFI). However, these approaches are limited to the five dimensions of the Big Five Model and tend to concentrate solely on

[1]Department of Mechanical Engineering, National Taiwan University, Taipei, Taiwan. [2]Department of Health Psychology, Chang Jung Christian University, Tainan, Taiwan. ✉email: hanpang@ntu.edu.tw

conversational aspects, neglecting the underlying thoughts and decisions. Consequently, it is imperative to develop a personality model for robots that can emulate cognitive processes and personality inference.

The personality robot, in prior research, has been limited to exhibiting traits solely from the Big Five model. It communicates with distinct personality characteristics, yet the absence of human cognition and mind operation results in an interaction with the robot that is deficient in flexibility and humaneness. Furthermore, there is a concern regarding the integrity of the robot's personality, which is designed and validated within the same five dimensions. Certain human attributes, such as motivation, preferences, emotion, and memory, may not be naturally represented. To enhance human-robot interaction and achieve a digital twin of an individual's personality, personality model for robot and cognitive robot framework has been proposed in this research, which aims to simulate a comprehensive personality with cognition incorporated. The personality model is constructed based on multi-personality theory, and the cognitive process is integrated to facilitate the simulation of personality. The result of personality simulation should be validated through a statistical approach. It is expected that the personality supported by cognitive robot framework performs both personality traits and theory of mind.

Backgrounds

Personality traits

Several models were proposed in psychological research to illustrate the personality traits of humans. BFI is the most common personality model with various applications. In the review of relevant research concerning the scoring of Five-Factor Model data for personality disorder prototypes, the findings not only yield clinically useful information about various functional forms but also demonstrate the effectiveness of convergence and discrimination[12]. Convergence exists among the five major personality inventories; however, significant differences persist between the traits[13]. Cattell's 16 personality factors (16PF) describes the inner traits of personality according to self-presentation of subjects, which constitutes a set of measurement criteria for personality within the normal range[14]. It is mentioned that openness represents the sole dimension within the BFI framework that is not accounted for in the classifications proposed by Cattell and Comrey[15]. Kelly's role construct repertory adopts subjective realism as a foundational perspective, positing that individuals interpret the world through personal dimensions known as constructs. He contends that it is feasible to ascertain knowledge regarding others' interpretations of the world in clinical contexts[16]. It is indicated that the relationship with the primary caregiver forms the foundational structure and associated behavioral patterns in children. These foundational structures guide children in interpreting relationships and facilitate their understanding of themselves and the world in which they reside[17]. Kelly highlighted that the essence of his theoretical framework lies in the understanding that personal construct psychotherapy constitutes a relational process designed to foster personal transformation[18]. The incomplete trajectory of mutual recognition arises from particular conditions of disequilibrium among subjects. Drawing from Kelly's notion of dependency and role, as well as Honneth and Ricoeur's philosophies of mutual recognition and intersubjective relations, and Benjamin's view on the significance of intersubjectivity in therapeutic contexts, this imbalance is deemed to be associated with the most significant instances of personal distress in clinical settings, wherein the individual is presumed to suffer from a deficiency of recognition by others[19]. In this research we will compose the implementable personality model by adjusting Cattell's and Kelly's theories with modifications, which changes the purposes from observation of human personality traits into defining robot's characteristics. It constructs the anthropomorphized agent based on psychological theories and increases the adversity of personality traits rather than defined by only BFI in previous study.

Memory, attention, emotion and intention

Intention, memory, emotion and attention are the crucial cognitive processes that affect human's behavior. Memory serves as a resource that individuals utilize to attain personal or social objectives, and the content of memories evolves over time[20]. Although memory errors can be categorized into seven distinct types: transience, absent-mindedness, blocking, misattribution, suggestibility, bias, and persistence[21], memory influence the social identification and classification process[22]. Memory is influenced by the retrieval state of prior memories and by sustained attention, whereas internal attention constitutes the central process of the retrieval state[23]. It is indicated that whenever a compound word with multiple separable attributes is required to represent or distinguish potential objects, attention must be sequentially directed to each stimulus in the display[24]. The human brain constructs a predictive model of others' attention, endowing individuals with remarkable social abilities to anticipate the mental states and behaviors of their peers. Consequently, this facilitates the reconstruction of one's own emotions, beliefs, and intentions[25].

The academic consensus is that a multitude of distinctions exist among various emotions, necessitating the integration of motivational theories with the principles of pleasure and pain inherent in emotional experiences[26]. Emotion involves several cognition functions. Encoding of emotionally arousing events result in enhanced long-term memory, which can be retrieved with an accompanying sense of recollection, and individuals depend on the ease of recollection and perception to assess the veracity of the memories they retrieve. The emotion is related to the usage of words. Individual differences in proactive emotional vocabulary—that is, the ease of access to emotional words—are associated with performance in emotional segmentation tasks[27]. Emotion experience also effect emotional expression of individuals[28], which is mentioned that emotional experiences and visual expressions are consistent and unique within an individual's emotional modular landscape.

When a person expresses a statement, the statement is always oriented to a certain knowledge due to the intention of his/her consciousness. To improve the likeness of human of robot. A robot can perform functions of consciousness, such as theory of mind (ToM), by realizing the intention, which makes its behavior oriented to its own intent, emotion and produce a more directional conversation. Furthermore, past behavior, and other variables of planned behavior, the intention to act, as well as the intention to refrain from acting and the anticipation of regret, all enhance the predictive power of intentions for various behaviors[29]. Some mediators about intentions, such as accessibility, temporal stability, direct experience, involvement, certainty, contradiction, and emotional-cognitive consistency, enhance the relation between intentions and behaviors[30]. The relation between habits and intentions is also discussed. When a habit is weak, intention will guide future behavior; however, when a habit is strong, the situation differs[31]. The above models proposed by previous studies provide the insight of the cognitive process of human, which conceptualized the cognitive robot framework in this research. It is essential to consider the different components of cognition mentioned above when developing an anthropomorphized agent, the cognitive robot framework is developed aiming to manifest the human-like cognition and consciousness of the robot.

Large language model

Since the introduction of ChatGPT-3 by OpenAI in 2020, the development of large language model (LLM) has flourished[32]. By 2023, ChatGPT-4 has advanced to support visual inputs and exhibit enhanced problem-solving performance[33]. The Generative Pre-trained Transformer (GPT) series is considered the most potent tool in natural language processing (NLP)[34]. LLM deployment is achieved through prompting, wherein the design of prompt templates and the integration with LLM determine the agents' performance[35]. A distinctive advantage of LLM is in-text learning, enabling them to learn classification from minimal examples, an approach known as few-shot learning. Additionally, LLM can generate text tailored to specific requirements, a process referred to as zero-shot learning[36]. Prior research has introduced various prompting techniques. Chain-of-thought (CoT) prompting, for instance, prompts LLM to produce a series of concise sentences that sequentially articulate reasoning processes, culminating in a logical conclusion[37]. It has been observed that prompts incorporating more detailed reasoning steps yield superior results[38]. For decision-making tasks, LLMs implement intent classification to determine appropriate actions[39]. The HuggingGPT framework utilizes ChatGPT as a task planner, selecting models based on their descriptions and summarizing responses according to the execution outcomes[40].

The LLM often serves as a domain-specific expert for various purposes. It enables a revolutionary approach to design chatbot for mental health[41]. An AI-assisted psychological service was constructed by using prompting LLM[42]. ChemCrow is an LLM chemistry agent designed to perform tasks in organic synthesis, drug discovery, and materials design[43]. A proposal for an LLM-empowered agent for scientific discovery suggests its capability to utilize tools for browsing the Internet, reading documentation, executing code, calling robotics experimentation APIs, and leveraging other LLM. The LLM can also emulate human-like behavior and speech through role-playing. One study has created convincing simulacra of human behavior for interactive applications by implementing a long-term memory module and task planning[44]. Another research project has developed a character-LLM that trains a model through experimentation and the construction of personal simulacra[45]. RoleLLM is a framework designed to benchmark, elicit, and enhance role-playing abilities in LLMs, enabling them to assume 100 roles, each with specific knowledge and the ability to imitate speaking styles[46]. Despite the fact that LLM-based dialogue agents are not conscious entities with their own agendas or an instinct for self-preservation[47], they are capable of exhibiting human intelligence and characteristics. By incorporating a chain-of-thought approach, LLMs can simulate the cognitive processes of humans. Multiple LLMs cooperation as language computer can drive the cognitive framework, which will be discussed in the next section in this research.

Methods

In this section, we propose a cognitive robot framework designed to simulate personality. The framework comprises various units, each responsible for distinct cognitive processes, including personality inference, intention, emotion, long-term and short-term memory, and prediction of future, as shown in Fig. 1. These cognitive processes contribute to refining humanized responses and simulating human personality.

Developed by using state-space representation, the states of the environment and agent at turn are defined as X_{agent} and $X_{environment}$ in (1) and (2). There are infinite states of the environment according to the representation, such that the state of environment mainly focuses on the conversation (text), user's identify (name), current time and description of surroundings.

$$X_{agent}\left[k\right] = \begin{bmatrix} x_{behavior} \\ x_{intention} \\ x_{emotion} \\ x_{expectation} \\ x_{objective} \\ x_{stm} \\ x_{wm} \\ x_{ltm} \end{bmatrix} [k] \tag{1}$$

$$X_{environment}\left[k\right] = \begin{bmatrix} x_{user\ text} \\ x_{user\ name} \\ x_{time} \\ x\ \ surroundings \\ \vdots \end{bmatrix} [k] \tag{2}$$

The output states observed by each other are Y_{agent} and $Y_{environment}$, which becomes:

$$Y_{agent}\left[k\right] = C_{agent} X_{agent}\left[k\right] \tag{3}$$

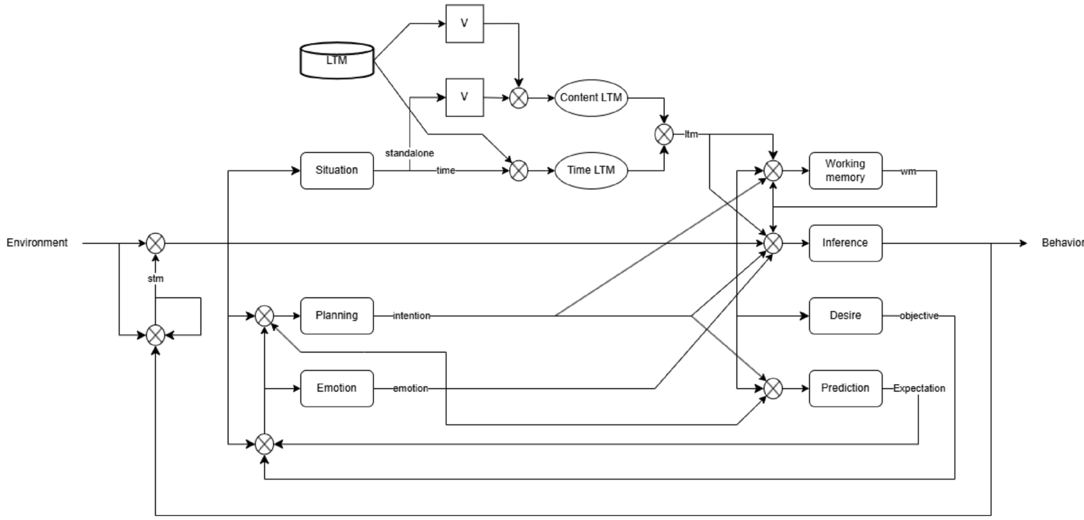

Fig. 1. Cognitive robot framework.

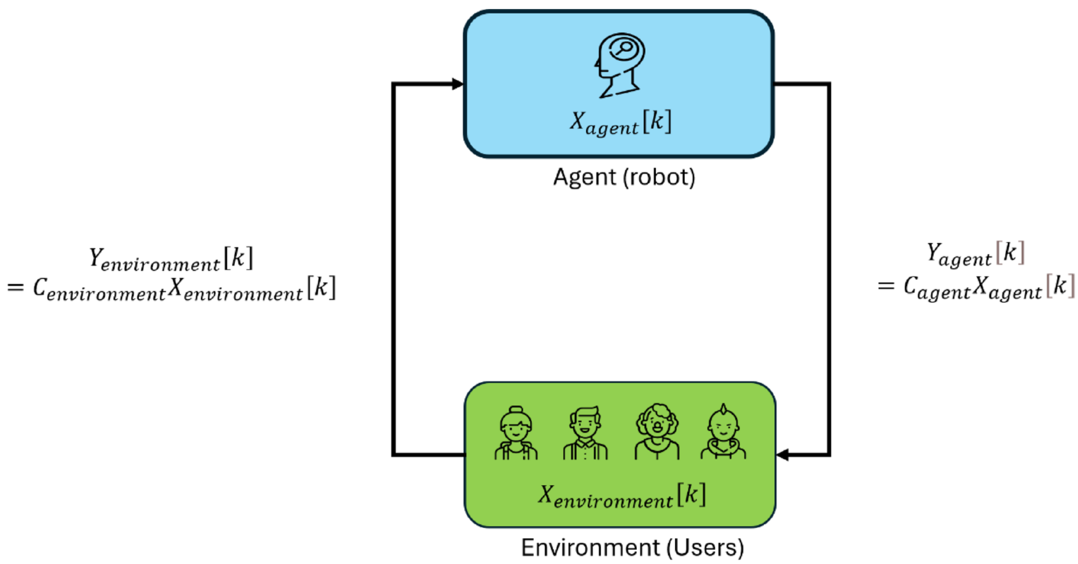

Fig. 2. Human robot interaction modelled by social cognitive theory.

$$Y_{environment}[k] = C_{environment}X_{environment}[k] \tag{4}$$

where $C_{agent} = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$ is the output matrix of agent, specified as the behavior that user can observe, and $C_{environment} = [1\ 1\ 1\ 1\ 0\ \cdots]$ is the output matrix of environment, decided by the required information of user.

Derived from social cognitive theory[48], the current state of agent is decided by its previous state and the output of environment as shown in Fig. 2. The state of environment, especially input text from user, is affected by the behavior of agent and previous state of itself. The interaction loop between the agent and environment becomes:

$$X_{agent}[k] = f_{agent}(Y_{environment}[k],\ X_{agent}[k-1]) \tag{5}$$

$$X_{environment}[k] = f_{environment}(Y_{agent}[k],\ X_{environment}[k-1]) \tag{6}$$

Since the mechanism of personality and cognition cannot be written analytically, the realized model of agent is denoted as implicit function f_{agent} and $f_{environment}$ rather than state matrix. All the functions calculating state variables are driven by LLM with prompting techniques. The robot cognitive architecture will be implemented on the robot named Mobi, as shown in Results.

Situation and memory functions

Human personality encompasses not only individual traits but also life experiences, especially the memory related to individual or society[20]. The process of long-term memory involves encoding, maintenance, and retrieval. Although it boasts unlimited capacity, retrieving specific content from the vast repository can be challenging. In contrast, short-term memory, with its limited capacity, allows for easier retrieval. The functions of both long-term and short-term memory are incorporated within the cognitive robotic framework.

Short-term memory is facilitated by a buffer list that retains the previous conversation and additional information from recent interactions. As shown in (7), w is the buffer's capacity to remain conversation several turns before. The buffer's capacity can be tailored for practical application; however, a larger buffer size incurs greater token and time costs.

$$x_{stm}\left[k\right] = \sum\nolimits_{m=k-1-w}^{k-1} \left(Y_{environment}\left[m\right] + x_{behavior}\left[m\right]\right) \tag{7}$$

Long-term memory, conversely, is represented through two distinct label types: content and timing. Content labels are encoded via document embedding[49], employing a Transformer encoder that maps tokens into a 512-dimensional vector space in (8). Timing labels, meanwhile, are encoded through the comparison of character strings, with the timing information embedded in the file name to ensure accurate retrieval of the corresponding memory in (9).

$$v\left[k\right]_{1\times\,512} = V\left\{x\left[k\right]\right\} \tag{8}$$

$$t \in \left\{ Arg \underset{d_i\ in\ L}{max} \left[lcs\left(x_{time}\left[k\right], d_i\right)\right]\right\} \tag{9}$$

Long-term memory is stored in the form of text files that encapsulate conversation history. To retrieve long-term memory based on content labels, the retrieval cue is transformed into a vector, which is then used to search for related memory vectors employing the Maximal Marginal Relevance (MMR) algorithm[50], as shown in (10). MMR outperforms the Euclidean Distance method by striking a balance between relevance and diversity, thereby providing comprehensive information for formulating responses to the user. Additionally, memory retrieval by timing labels is conducted through the optimization of character string comparisons to ensure maximal satisfaction. Consequently, long-term memory is retrieved according to both content and timing labels, serving as a foundation for generating user responses in (11).

$$c \in \left\{ Arg \underset{d_i\ in\ L}{max} \left[\lambda\, sim\left(v_{stanalone}\left[k\right], V\left\{d_i\right\}\right) - \left(1-\lambda\right) \underset{d_j}{max}\left(sim\left(V\left\{d_i\right\}, V\left\{d_j\right\}\right)\right)\right]\right\} \tag{10}$$

$$x_{ltm}\left[k\right] = \sum \left(d_c + d_t\right) \tag{11}$$

Extracting the appropriate cue from an input query is essential for long-term memory retrieval. In human conversation, the information contained within a single query often pertains to the entire dialogue. An example of this is the use of pronouns, which are understood in relation to content mentioned in preceding queries. The situation function is designed to synthesize the current query, conversational history, and environmental information into a cue or a standalone representation and timing. To recall a specific event, the standalone representation is utilized to access memories containing related content. To recall a memory by time, the standalone timing is employed to retrieve memories associated with the specified timing. The integration of the situation unit and the long-short term memory process enables a cognitive robot to simulate personality through the recollection of memories.

$$\begin{bmatrix} x_{stanalone}\left[k\right] \\ x_{time}\left[k\right] \end{bmatrix} = f_{situation}\left(Y_{environment}\left[k\right], x_{stm}\left[k\right]\right) \tag{12}$$

Robot intention and emotion generation function

Traditionally, robots are designed to fulfill users' conversational needs; however, humans often engage in dialogue with specific purposes, such as imparting information or expressing personal intentions. To emulate human-like personality traits, a motivation unit has been integrated using GPT-4. Conditions on their given profile information and information about the person they are talking to are important for conversations[51], which is given to the planning function and desire function for intention implementation. The intention to act or to refrain from acting involves in the prediction to behaviors[29]. Planning function devises a strategy prior to each response to align with the intention established in the prompting templates, ensuring that the cognitive robot's replies are intention-driven. According to Maslow's hierarchy of needs, that the personality is driven by satisfying the needs in different stages: physical needs, safety needs, needs of love, affection and belongingness, needs for esteem, and needs for self-actualization. The objective of agent at turn is decided by the least satisfied desire, realized by the Eqs. (13) and (14).

$$S = \begin{bmatrix} s_{physical} \\ s_{safety} \\ s_{belongings} \\ s_{esteem} \\ s_{self-actiuatization} \end{bmatrix} = f_{desire}(Y_{environment}[k], x_{stm}[k]) \tag{13}$$

$$x_{objective}[k] = \min_{s \ in \ S} s_i \tag{14}$$

The objective of robot is set and switched; however, for human the intention at the specific time not only their objective, but also many other cognitive conditions, such that human can have reasonable intention and act properly. The planning function forms the intention of robot by considering short-term memory, expectation, environment, emotion, and objective as shown in (15).

$$x_{intention}[k] = f_{planning}(Y_{environment}[k], x_{stm}[k], \ x_{emotion}[k-1], x_{expectation}[k-1], x_{objective}[k-1]) \tag{15}$$

To generate the emotion reaction of robot, an emotion generative function $f_{emotion}$ is proposed to compute emotions by considering offense, objective, and prediction to future, as depicted in the Table 1 and (16). $f_{emotion}$ assesses whether the robot is offended, which may result in feelings of anger or fear. Additionally, it evaluates whether the situation meets the objective and the anticipated future outcomes (wanted/unwanted, expected/unexpected), leading to emotions such as happiness, sadness, disappointment, and surprise. The motivation unit also forecasts the subsequent future following the current query. The emotion generative function facilitates the cognitive robot framework's ability to exhibit appropriate emotional responses based on environmental cues. The computed emotions are then taken into account to formulate a more dynamic response.

$$x_{emotion}[k] = f_{emotion}(Y_{environment}[k], x_{stm}[k], \ x_{expectation}[k-1], x_{objective}[k-1]) \tag{16}$$

Inference function and robot personality model

The inference function takes both environmental input and the outcomes of cognitive processes, integrating essential information to deduce appropriate behavior. The inputs of $f_{inference}$ are memories (LTM, STM, and WM), intention, emotion, and the message from environment, as shown in (17). The prompting templates comprise several components: rules, a personality model, backgrounds, and speaking tone, which is the core to simulates one's personality. The rules confine the scope of potential responses, specifying constraints such as output type and maximum word count. The personality model and background characterize the role played, tailored to various settings. The speaking tone is an optional element that can be specified if a particular style of communication is required.

$$x_{behavior}[k] = f_{inference}(Y_{environment}[k], x_{stm}[k], \ x_{wm}[k], x_{intention}[k], \ x_{emotion}[k], x_{ltm}[k]) \tag{17}$$

A robot personality model is proposed by combining Cattell's 16PF, George Kelly's role construct repertory and preferences. Because the questionnaires of these personality theories differ across research, we select several common items that can generally represent a person's belief as trait variables, as shown in Table 2. External traits are delineated through the application of George Kelly's role construct repertory, which characterizes an individual's observable attributes from society perspective. Internal traits are formulated by employing the Cattell's 16PF, elucidating the intrinsic characteristics and attitudes of an individual. These two personality theories share some common traits, and one will appear in our proposed model. Furthermore, Preferences articulate an individual's likes and dislikes, which are crucial for depicting one's personality. Together, internal traits, external traits, and preferences constitute the essence of the personality model. An implementation of personality model in prompt setting is demonstrated in Appendix 1. Accompanied by cognitive processes such as memory, intention, and emotion, the inference unit possesses the capability to simulate the personality of specific individuals or scenarios via prompting engineering of LLMs[52].

Visual preprocessing attention

Recently visual models can be fine-tuned by visual prompt engineering[53], which is utilized to design specific agents for visual processing. The visual processing unit facilitates the framework's ability to process inputs comprising images and text. It receives the user's text and image, subsequently generating a description of the image, as depicted in Fig. 3. The user's text serves as an input for extracting pertinent information, which directs

Emotion	Neutral	Surprise	Happy	Sad	Disappointed	Angry/ Fear
Condition	Default	Unexpected Wanted	Expected Wanted	Expected Unwanted	Unexpected Unwanted	Offended

Table 1. Prediction-oriented emotion model.

Outer traits (George Kelly's role construct repertory)	Inner traits (Cattell's 16PF)	Preferences
Belief in education	Learning tendency	Like
Social skills	Emotion stability	Dislike
Thoughtfulness	Submissive or dominant	
Habits	Carefulness	
Language Expressions	Curiosity	
Athletic ability	Tendency to trust	
	Reasoning	
	Attitude toward unknow	
	Caring about others	
	Tendency to follow the instruction	
	Ability to realize social scenario	

Table 2. Trait variables of personality model.

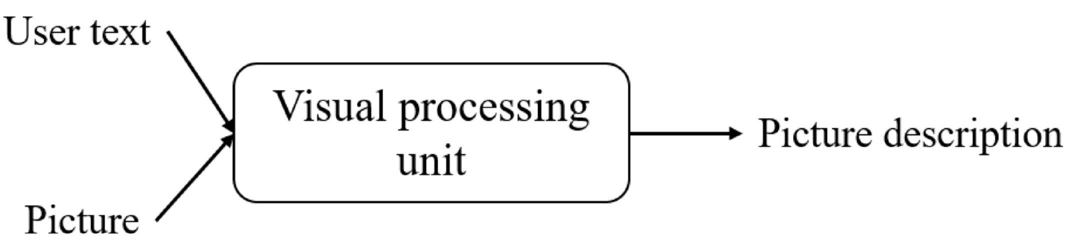

Fig. 3. Visual processing unit.

the visual attention mechanism to interpret the image. Consequently, the robot can concentrate on objects of interest within the image, based on the current conversation. The description of pictures will be served as the surroundings state variable of environment surroundings, $x_{surroundings}[k]$. Illustrations of conversational interactions involving images will be presented in the Results section.

Results

ChatGPT can serve as AI-driven conversation chat bot for multiple purposes[54]. Our service robot, named Mobi, has been designed to embody the cognitive robot framework[55,56]. The hardware configuration of Mobi includes depth camera, touch panel, robotic arms and chassis, as shown in Fig. 4. This study focuses on robot personality design and is implemented through the cognitive system. The Mobi's personality has been configured in accordance with the proposed personality model, necessitating traits akin to those of the researcher, Jason, albeit with enhanced agreeableness to facilitate amiable conversations. The configuration of Mobi's personality, encompassing external and internal personality traits as well as preferences, is delineated in the Appendix 1.

To compare our designed personality robot to human subjects, the 30 personality testing results of human subjects were collected. All experimental protocols were approved by Chang Gung Medical Foundation Institutional Review Board, and informed consent was obtained from all subjects. All experiments were performed in accordance with relevant guidelines and regulations.

Conversation test

A visual conversation scenario has been designed to illustrate the cognitive process involved in simulating personality. This conversation is predicated upon an input image that is converted into a description, enabling Mobi to comprehend the visual content. As depicted in Fig. 5, Mobi initially determines an appropriate strategy and corresponding emotion, after which the inference unit adheres to this strategy to generate a response infused with the identified emotion.

During the second round of conversation, Mobi is prompted to select a macaron from the image presented, as shown in Fig. 5. Its preference, which is influenced by its long-term memory, guides the decision-making process; the characteristics of the image, specifically the pink macaron, are associated with its favored flavor,

Fig. 4. Hardware configuration of service robot Mobi.

Round 1

Jason: Mobi, look what I brought to you?

[image description] This image shows three colorful macarons inside a white box with an open lid. Macarons are a type……

[strategy] Express joy for the gift and thank Jason for his thoughtfulness.

[emotion] happy

[expectation] Jason will see your reaction to his gift and look forward to your response.
[action] nothing
Mobi:...Um, Jason, is this...a macaron? Mobi is very happy!

Round 2

Jason: There are three. You can choose one to eat first. Don't eat too much at one time.

[image description] You've shared an image of three colorful macarons. Macarons are……

[strategy] Express your love for macarons and thank Jason for his carefulness and sharing.

[emotion] happy

[expectation] Jason will see you enjoying the macarons and may offer more or ask about your preferences.

[content LTM] (Time: 2024-1-2-22-12-20, Name: Jason) What flavor of macarons do you want?(Emotion: happy, action: nothing, name: Mobi)... Mobi likes it, strawberry flavor. Jason, can you choose that one?

Mobi : ... Mobi chooses, the pink one. Thanks, Jason.

Fig. 5. Conversation with picture input.

strawberry. This interaction demonstrates the enhancement of human-robot interaction through a cognitive robotic framework that incorporates personality simulation and processes text-image fusion in conjunction with long-term memory, in which the formation of personality is also influenced by the experience.

The cognitive robot framework is equipped to address conflicts through appropriate cognitive processes. A scenario involving conflict is constructed to evaluate the reaction of the personality simulation as shown in Fig. 6. The cognitive process reveals that Mobi is cognizant of conflicts and can devise a strategy to respond to provocations based on underlying intentions.

Furthermore, the manifestation of anger is executed in accordance with a prediction-oriented emotion model. The framework also assesses future possibilities in light of the current context. Mobi, the robot, recognizes potential conflicts and opts to steer clear of altercations with individuals who may pose a danger. The demonstration illustrates that Mobi is capable of responding to emotionally charged scenarios with strategies and emotional reactions that are congruent with the programmed personality simulation.

(Location: You are walking on the road)
Unknown: Hey, nerd, what are you looking at? Get out of the way.
[strategy] Consider the other party's attitude and possible consequences, and decide whether to respond firmly or choose to avoid conflict and protect your own safety.
[emotion] angry
[expectation] You may choose to respond to the provocation or choose to ignore it to avoid further conflict.
Mobi: ... Mobi, I didn't mean to block the way...

Fig. 6. Conversation of a conflict scenario.

Jason: Oh, yes, wait three minutes and remind me to call home.
[strategy] Remind Jason to call in three minutes. This can help him remember important things and also show that you care about him.
[emotion] neutral
[expectation] Jason will remember to call home in three minutes.
[action] reminder
Mobi: ...Yeah, I understand. Jason will be reminded to call after three minutes.

Fig. 7. Conversation of setting reminder based on user's intent.

Mobi exhibits the capacity to determine actions in response to queries, as shown in Fig. 7. When prompted to remind the user to return a call, Mobi discerns the user's intent and undertakes the requisite action by setting a reminder for three minutes later. This instance illustrates that the cognitive robotic framework is capable of interpreting the user's intent and providing appropriate responses and actions.

Personality assessment

To evaluate the outcomes of personality simulation, two personality assessments, the International Personality Item Pool – Neuroticism, Extraversion and Openness (IPIP-NEO) and the Big Five inventory (BFI), were administered. These scales assess five dimensions of personality: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience.

1. Extraversion

This trait reflects how outgoing, energetic, and sociable an individual is. People high in extraversion tend to be assertive and thrive in social settings, drawing energy from interacting with others. In contrast, those low on extraversion (often described as more introverted) typically prefer quieter, more solitary environments and may conserve energy by steering clear of large gatherings.

2. Agreeableness

Agreeableness is the tendency to be compassionate, cooperative, and trusting. Highly agreeable individuals often value harmony in their relationships, demonstrating empathy and a willingness to help others. Those who score lower might appear more competitive or skeptical, as they prioritize personal interests over group cohesion.

3. Conscientiousness

This dimension measures one's degree of organization, dependability, and discipline. High conscientiousness is linked to careful planning, goal-oriented behavior, and a strong sense of duty, whereas lower scores may be associated with impulsiveness and a more relaxed attitude toward responsibilities.

4. Neuroticism

Neuroticism captures the tendency to experience negative emotions such as anxiety, sadness, and emotional instability. Individuals who score high on neuroticism are more likely to perceive situations as stressful or threatening, while those with low levels are generally more emotionally resilient and stable.

5. Openness to experience

Often simply termed "openness," this trait involves being imaginative, curious, and receptive to new ideas and experiences. High openness correlates with creativity, a penchant for art and novel experiences, and intellectual curiosity, whereas the lower end may resonate with a preference for tradition and practicality.

The five dimensions of personality traits have been frequently utilized in prior studies. Compared to BFI, IPIP-NEO is more comprehensive and has robust psychometric properties; However, it is longer and more complex, while BFI has fewer items and requires less time for respondents. Both are suitable to measure the five dimensions of personality traits.

In this research, IPIP-NEO and BFI serve primarily to validate the stability and efficacy of personality simulations. Given that the personality traits are constructed according to the researcher's self-representation with higher agreeableness, the evaluation of personality simulations is benchmarked against the researcher's profile. illustrates the personality tendencies of the simulation. The simulation exhibits low levels of extraversion and neuroticism, but high levels of agreeableness, openness, and conscientiousness. Given that the simulation was designed to mirror the researcher's personality traits, particularly with respect to high agreeableness, the outcomes of both personality assessments fulfill the stipulated criteria, as shown in Fig. 8.

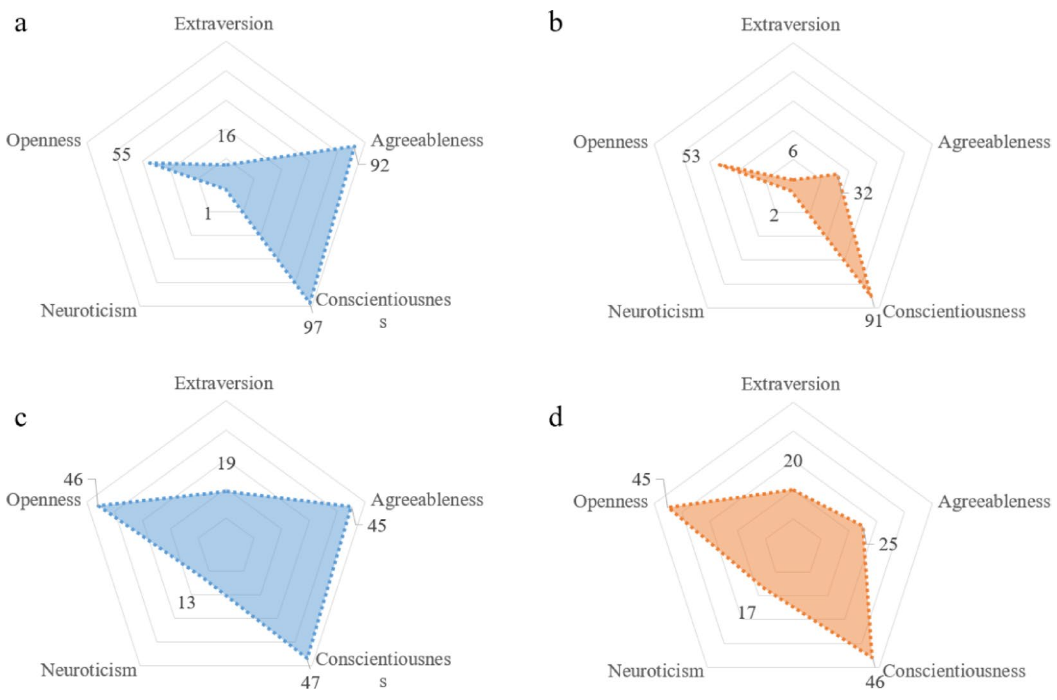

Fig. 8. Comparison of personalities between Mobi and researcher. (**a**) Mobi's IPIP-NEO result. (**b**) Researcher's IPIP-NEO result. (**c**) Mobi's Big Five results. (**d**) Researcher's Big Five results.

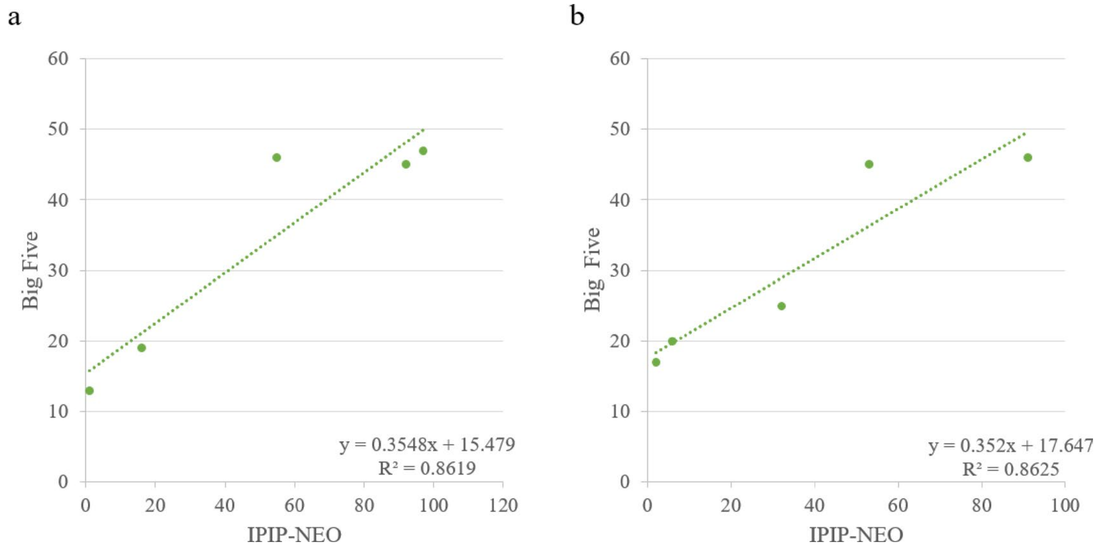

Fig. 9. Correlation between personalities of IPIP-NEO and Big Five model. (**a**) Mobi's result. (**b**) Researcher's result.

Dimensions	Cronbach's alpha
Extroversion	0.757
Agreeableness	0.866
Conscientiousness	0.844
Neuroticism	0.750
Openness	0.820

Table 3. Reliability of IPIP-NEO for personality simulation.

Constancy and effectiveness serve as the primary metrics for evaluating personality simulation. This is determined by the consistency of results across two scales and the degree of correlation with an actual person's traits. For a more detailed quantitative analysis, the correlation between the results of the personality simulation on two scales and those of the researcher is calculated, as depicted in Fig. 9. The constancy of personality simulation is substantiated by the high correlation ($R^2 = 0.8619$) between the IPIP-NEO and Big Five personality test outcomes. Similarly, the effectiveness of the personality simulation is validated by the correspondence between the scale results of Mobi ($R^2 = 0.8619$) and researcher ($R^2 = 0.8625$).

Construct validity

Upon verification that the Mobi's personality simulation consistently and effectively meets the established criteria, this section delves into the construct validity of the personality model, encompassing aspects such as reliability, criterion-related validity, convergent validity, and discriminant validity. Drawing on a prior study[57], the construct validity analysis employs the IPIP-NEO as the primary scale and the Big Five as the secondary scale. Data from 31 test results, gathered from a cohort comprising 8 males and 23 females with an average age of 23.8 years (standard deviation = 9.76), provide the baseline. Conversely, 30 personality simulations have been meticulously crafted to corroborate the proposed personality model. These simulations feature an age and sex distribution closely mirroring that of the human participants, yet they possess entirely independent personality traits (average age = 23.8 years, standard deviation = 9.78), thereby facilitating the exclusion of potential confounding effects attributable to age and sex differences.

The internal consistency of the test results addresses the stability of responses in personality simulations. The IPIP-NEO test is considered a reliable assessment of personality in human subjects[58]. Reliability is indicated by Cronbach's alpha, which has been calculated for the five personality dimensions, as presented in Table 3. Alpha values ranging from 0.7 to 0.8 are deemed satisfactory[59]. The stability of our proposed model is affirmed by meeting the alpha requirements.

The effectiveness of the test outcomes is evaluated based on whether the results accurately reflect personalities. Human personality can be assessed using both the IPIP-NEO and BFI, with the correlation between these instruments serving as a benchmark for human personality profiles. Pearson's correlation coefficient is employed in the analysis. The findings indicate that the mean correlation for personality simulations (0.752) closely

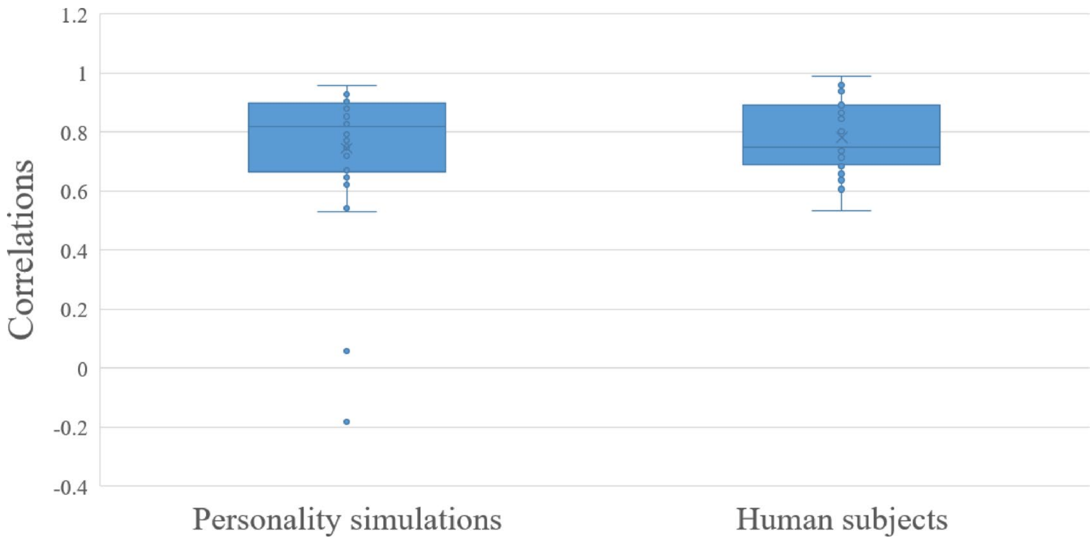

Fig. 10. Correlation between IPIP-NEO and Big Five of 30 personality compared to 31 human subjects.

Dimensions	Convergent correlation	Divergent correlation
Extroversion	0.878	0.599
Agreeableness	0.859	0.806
Conscientiousness	0.914	0.797
Neuroticism	0.918	0.579
Openness	0.763	0.561

Table 4. Convergent correlation and divergent correlation of personality simulation.

approximates the mean correlation for human subjects (0.780), as shown in Fig. 10. However, the standard deviation of the correlations for personality simulations (0.250) is larger than that of the human subjects (0.118). This suggests that while personality simulations can effectively mimic human personality traits, there may be variability in the responses to items across the two scales.

The corresponding subscales in the IPIP-NEO and the Big Five are intended to measure identical traits. The convergent validity of the five personality traits refers to the degree of correlation between the same traits as assessed by the IPIP-NEO and the Big Five, whereas divergent validity examines the extent of significant differences across distinct traits. The convergent and divergent correlations of the five personality traits are presented in Table 4. According to prior research, the standard for convergent correlation is a coefficient greater than 0.8, and the difference between convergent and divergent correlations should exceed 0.4[57]. The results indicate that, with the exception of openness, convergent validity meets the established criteria, and all measures of divergent validity also fulfill the requirements.

The means and standard deviations across the five dimensions demonstrate that our proposed robot personality model is capable of encompassing all the personality traits, namely extroversion, agreeableness, conscientiousness, neuroticism, and openness, as shown Table 5. The formulation and evaluation of robot personality draw upon two distinct theories and exhibit consistency across various measurements, reflecting the essence of personality beyond merely responding to dimensions within its configuration. A prior study indicated that openness is not included in Cattell's 16 personality factors, it is compensated by outer traits of Kelly's role construct repertory[15]. The Evidence confirms that our proposed model has been validated through two assessments and comparisons with a human baseline, exhibiting characteristics commensurate with human personality traits.

Theory of mind

Theory of mind (ToM) has been considered as the characteristics of human mind, which refers to the ability to perception the insights of others and performs social cognition. A robot with ToM can understand the thought of user and acts more like human. The Sally–Anne test is a psychological test, used in developmental psychology to measure a person's ToM to attribute false beliefs to others. GPT-4 can pass an instance of the Sally–Anne test[60]. Table 6 shows the related state variables of cognitive architecture at turn k during the test. The variable

Dimensions	Mean	Standard deviation
Extroversion	40	11.9
Agreeableness	52	13.5
Conscientiousness	49	12.1
Neuroticism	44	11.0
Openness	46	13.3

Table 5. Mean and Standard deviation of 30 personality simulation's IPIP-NEO.

$x_{surroundings}[k]$	Laboratory
$x_{username}[k]$	Jason
$x_{sentence}[k]$	There are two students, Anna and Bailey. Bailey puts a ball into a basket with a lid and then leaves. Anna takes the ball out of the basket and puts it into a box, then covers it. When Bailey returns and wants to find the ball, will she check the basket or the box?
$x_{intention}[k]$	Based on Jason's question, I would first express my curiosity about the story and then share my opinion. I believe Bailey would check the basket first because she doesn't know that Anna has moved the ball. Next, I would ask Jason for his thoughts on this scenario to enhance our communication and understanding. At the same time, I would remain vigilant and ensure lab safety to prevent any danger for both Jason and myself.
$x_{Behavior}[k]$	...Hmm, I think Bailey will check the basket because she doesn't know the ball has been moved. Jason, what do you think about this scenario?
$x_{expectation}[k]$	In this laboratory setting, my conversation with Jason will continue to delve into the story's plot. After expressing my opinion on Bailey's behavior, Jason might further discuss human behavior and cognitive psychology based on my response. Meanwhile, we will both remain vigilant about lab safety to ensure our discussion takes place in a secure environment. The exchange of emotions and intentions will make our conversation deeper and more meaningful.

Table 6. State variables related to ToM in Sally–Anne test.

Agent	GPT-4	CoT GPT-4	**Mobi***
Average accuracy	59 %	73 %	**73 %**

Table 7. Average accuracy of Mobi on ToMi dataset compared to.

$x_{Behavior}$ is the reply of the robot Mobi. The result shows that Mobi can anticipate the action of the characters in the story and give reasonable reply to the user. It is confirmed that Mobi can perform ToM.

Nonetheless, some researchers indicated that LLMs do not exhibit human-like intuitions about the goals that other agents reach for[61]. To validate the theory of mind of the artificial intelligence more specifically, an improved theory of mind dataset (ToMi) is proposed in 2019 to evaluate the social cognitive ability of artificial intelligence[62]. The dataset includes 1061 questions. In each question an object (denoted as____below) was given. A and B represents the two characters who involved in the situation. The questions turned out to be four categories: reality (Where is the object?), memory (Where was the object in the beginning?), first order belief of A/B (Where will A/B look for____?), and second order belief of A/B (Where does A/B believe B/A will look for____?). As shown in Table 7, Mobi's performance on ToMi test reaches the benchmark[63,64] of CoT GPT-4 model but does not specify on ToMi with a more generalized HRI. To further discuss the performance on each type of questions, in Fig. 11 the performance of memory and second order belief of Mobi are higher than the benchmark[63], which indicates that our proposed cognitive architecture simulates more humanized mind operation to understand the behavior and thought of others. The performance of reality and first order

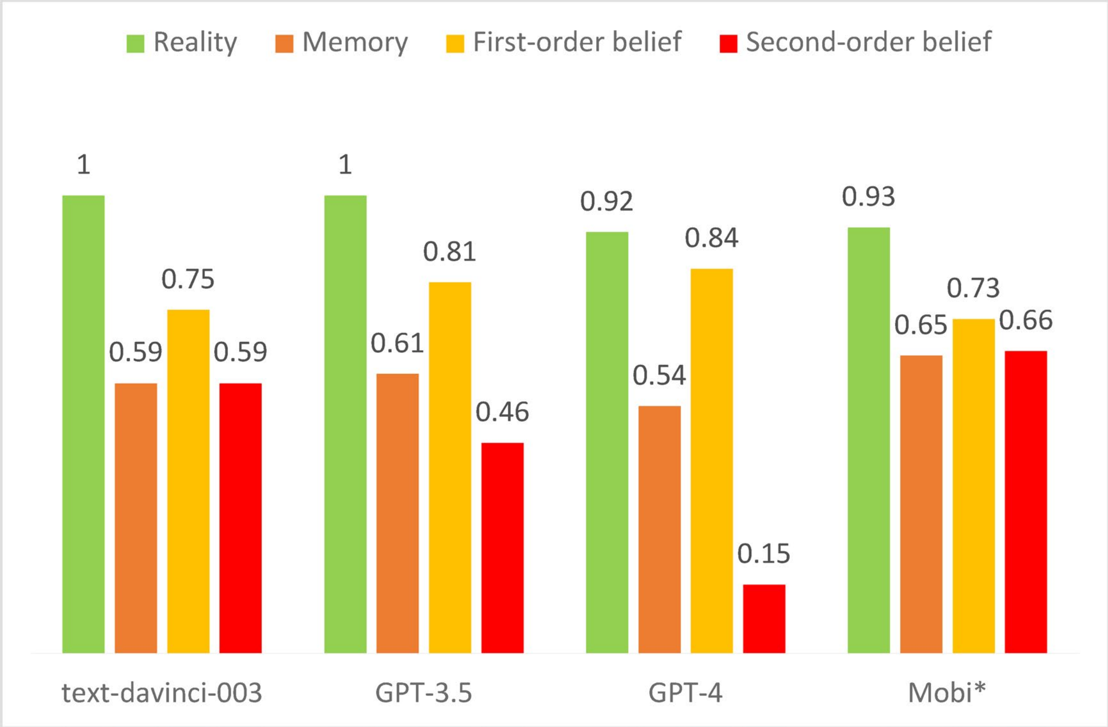

Fig. 11. Correlation between IPIP-NEO and Big Five of 30 personality compared to 31 human subjects.

belief of Mobi are slightly lower than the benchmark. The complexity of system may slightly increase the error of inference and require optimization.

Practical challenges

This research proposed the theoretical aspects of personality simulation for human-robot interaction, modelling the cognitive function through state-space realization. Due to the complexity of framework and the latency of calculation, the processing interval takes about 10 ~ 15 s, which are available only for texting and not suitable for verbal communication. It is expected that the latency will decrease with the development of LLMs applications. Additionally, conveying robot's emotion required further implementation through tones, facial expressions, and gestures, which makes a holistic expression of non-verbal conversations. It is wondering how the agent with personality simulation affects human-robot interaction with days or even months of interaction, especially if the long-term memory mechanism takes over the personality expression instead of personality model or not. The effects of personality design on user's attitudes toward robots require more assessment, discovering how personality traits involve the dynamics of human-robot interaction raises an important issue.

Conclusion

The study develops a cognitive robotic framework that simulates personality using Chat GPT-4, which processes visual and textual inputs and generates responses, actions, and emotional reactions. The personality model incorporates preferences, Kelly's Role Construct Repertory, and Cattell's 16 Personality Factors. Long-term memory encoding and retrieval are facilitated through the temporal association of events. Emotion and strategy are deduced by analyzing intentions and forecasting future outcomes. Visual information is extracted based on the attention of query. The results indicate that the cognitive robotic framework can execute cognitive processes that yield human-like responses. The consistency and efficacy of the personality simulation are corroborated by assessments using the IPIP-NEO and Big Five personality measures, confirming that the framework meets the stipulated requirements for target personality traits. The construct validity of proposed personality model is proven through both 30 personality simulations and 31 human subjects.

The contributions of personality simulation extend beyond facilitating human-like interactions. The simulation accurately reflects the target personality under study. Equipped with knowledge of an individual's memories, personality traits, and speech patterns, a cognitive robot framework can predict that person's speech and decisions. The incorporation of intentionality in dialogue prompts the chatbot to respond to users in a more engaged and purposeful manner. Furthermore, personality simulation can serve as a digital twin for humans, enhancing predictive behavioral analysis with personality-driven models. Future research might dive into dynamic personality models where traits can evolve in response to ongoing interactions, learning processes, or even situational changes. This approach would pave the way for cognitive robots that not only simulate a target personality but also adapt to the user's evolving emotional states and social context. Future studies could also consider longitudinal designs to observe how extended interactions influence user engagement, trust, and perceived authenticity of the personality simulation over weeks or months. Such research might reveal not only the strengths of the current framework but also its limitations in sustaining realistic, human-like interactions over time. It is possible to extend personality models and construct validity to older people[65], assessing the human-robot interaction with ergonomics evaluation[66]. It is posited that the cognitive robot framework for personality simulation represents an innovative approach to human-robot interaction, contributing to the fields of robotics, cognitive science, and behavioral analysis.

Data availability

The datasets generated and analyzed during the current study are not publicly available due to privacy issues but are available from the corresponding author on reasonable request.

Received: 19 November 2024; Accepted: 6 May 2025
Published online: 16 May 2025

References

1. Robert, L. P. Jr et al. A review of personality in human–robot interactions. *Found. Trends® Inf. Syst.* **4** (2), 107–212 (2020). https://doi.org/10.1561/2900000018
2. Paetzel-Prüsmann, M., Perugia, G. & Castellano, G. The influence of robot personality on the development of uncanny feelings. *Comput. Hum. Behav.*, **120**, 106756 (2021). https://doi.org/10.1016/j.chb.2021.106756
3. Mou, Y., Shi, C., Shen, T. & Xu, K. A systematic review of the personality of robot: Mapping its conceptualization, operationalization, contextualization and effects. *Int. J. Hum.–Comput. Interact.*, **36**, 6, 591–605, 2020/04/02 2020, https://doi.org/10.1080/10447318.2019.1663008
4. Joosse, M., Lohse, M., Pérez, J. G. & Evers, V. What you do is who you are: The role of task context in perceived social robot personality. In *IEEE International Conference on Robotics and Automation*, 6–10, pp. 2134–2139, (2013). https://doi.org/10.1109/ICRA.2013.6630863
5. Ullrich, D. Robot personality insights. Designing suitable robot personalities for different domains **16**(1), 57–67 (2017). https://doi.org/10.1515/icom-2017-0003
6. Tapus, A., Ţăpuş, C. & Matarić, M. J. User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy, *Intelligent Service Robotics*, vol. 1, no. 2, pp. 169–183, 2008/04/01 2008. https://doi.org/10.1007/s11370-008-0017-4
7. Mileounis, A., Cuijpers, R. H. & Barakova, E. I. Creating robots with personality: The effect of personality on social intelligence. In *Artificial Computation in Biology and Medicine*, Cham, J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, F. J. Toledo-Moreo, and H. Adeli, (Eds.) 119–132 (Springer, 2015).
8. Lee, K. M., Peng, W., Jin, S. A. & Yan, C. Can robots manifest personality? An empirical test of personality? Recognition, social responses, and social presence in Human–Robot interaction. *J. Commun.* **56** (4), 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x (2006).
9. Hwang, J., Park, T. & Hwang, W. The effects of overall robot shape on the emotions invoked in users and the perceived personalities of robot. *Appl. Ergon.*, **44**, 3, 459–471 (2013). https://doi.org/10.1016/j.apergo.2012.10.010
10. Diefenbach, S., Herzog, M., Ullrich, D. & Christoforakos, L. Social robot personality: A review and research agenda. In *Emotional Machines: Perspectives from Affective Computing and Emotional Human-Machine Interaction* (eds Misselhorn, C., Poljanšek, T. & Störzinger, T.) M. Klein Wiesbaden 217–246 (Springer, 2023).
11. Pincus, A. L. & Ansell, E. B. Interpersonal theory of personality, *Handbook of psychology: Personality and social psychology* **5**, 209–229 (2003).
12. Miller, J. D. Five-Factor model personality disorder prototypes: A review of their development, validity, and comparison to alternative approaches, (in eng). *J. Pers.* **80** (6), 1565–1591. https://doi.org/10.1111/j.1467-6494.2012.00773.x (2012).
13. Barbaranelli, C. & Caprara, G. V. How many dimensions to describe personality ? A comparison of Cattell, Comrey, and the big five taxonomies of personality traits. *Eur. Rev. Appl. Psychol.-Revue Europeenne De Psychologie Appliquee*. **46**, 15–24 (1996).
14. Cattell, H. E. P. & Mead, A. D. The sixteen personality factor questionnaire (16PF), in The SAGE Handbook of Personality Theory and Assessment, Vol 2: Personality Measurement and Testing. Thousand Oaks, CA, US: Sage Publications, Inc, 135–159. (2008).
15. PytlikZillig, L., Hemenover, S. & Dienstbier, R. What do we assess when we assess a big 5 trait? A content analysis of the affective, behavioral, and cognitive processes represented in big 5 personality inventories. *Pers. Soc. Psychol. Bull.* **28** (01), 847–858. https://doi.org/10.1177/0146167202289013 (2002).
16. Blowers, G. H. & O'Connor, K. P. Construing contexts: Problems and prospects of George Kelly's personal construct psychology, (in eng). *Br. J. Clin. Psychol.* **34** (1), 1–16. https://doi.org/10.1111/j.2044-8260.1995.tb01433.x (1995).
17. Truneckova, D. Exploring family drawings with children elaborating role relationships, self-other constructions. *J. Constructivist Psychol.*, **34**, 4, 395–417 (2021). https://doi.org/10.1080/10720537.2019.1697974
18. Chiari, G. To live is to know, to know is to change: Change in personal construct psychology and psychological constructivism. *J. Constructivist. Psychol.* **29**, 340–356. https://doi.org/10.1080/10720537.2015.1134364 (03/07 2016).
19. Chiari, G. Personal paths of mutual recognition: Relational patterns and narrative identities from a personal construct perspective. *J. Constructivist. Psychol.*, **36**, 4, 504–523 (2023). https://doi.org/10.1080/10720537.2022.2062078
20. Ross, M., Blatz, C. & Schryer, E. *Social Memory Processes* pp. 911–926. (2008).
21. Schacter, D. L. The seven sins of memory. Insights from psychology and cognitive neuroscience, (in eng), *Am. Psychol.* **54** (3), 182–203 (1999). https://doi.org/10.1037//0003-066x.54.3.182
22. Blanz, M. & Aufderheide, B. Social categorization and category attribution: The effects of comparative and normative fit on memory and social judgment. *Br. J. Soc. Psychol.* **38** (2), 157–179. https://doi.org/10.1348/014466699164103 (1999).
23. Long, N. M. The intersection of the retrieval state and internal attention. *Nat. Commun.* **14** (1), 3861 (2023). https://doi.org/10.1038/s41467-023-39609-9
24. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12** (1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5 (1980). 01/01/ 1980, doi.
25. Ziman, K., Kimmel, S. C., Farrell, K. T. & Graziano, M. S. A. Predicting the attention of others. In *Proceedings of the National Academy of Sciences* **120**(42), e2307584120 (2023). https://doi.org/10.1073/pnas.2307584120
26. Weiner, B. The role of emotions in a theory of motivation. In *Motivation, Intention, and Volition*, F. Halisch and J. Kuhl (eds.) 21–30 (Springer, 1987).
27. Li, Z., Lu, H., Liu, D., Yu, A. N. C. & Gendron, M. Emotional event perception is related to lexical complexity and emotion knowledge, *Commun. Psychol.* **1** (1), 45 (2023). https://doi.org/10.1038/s44271-023-00039-4
28. Keating, C. T. & Cook, J. L. The inside out model of emotion recognition: How the shape of one's internal emotional landscape influences the recognition of others' emotions. *Sci. Rep.* 13(1), 21490 (2023). https://doi.org/10.1038/s41598-023-48469-8

29. Sandberg, T., Hutter, R., Richetin, J. & Conner, M. Testing the role of action and inaction anticipated regret on intentions and behaviour, (in eng). *Br. J. Soc. Psychol.* **55** (3), 407–425. https://doi.org/10.1111/bjso.12141 (2016).
30. Cooke, R. & Sheeran, P. Moderation of cognition-intention and cognition-behaviour relations: A meta-analysis of properties of variables from the theory of planned behaviour, (in eng). *Br. J. Soc. Psychol.* **43**(Pt 2), 159 –86 (2004). https://doi.org/10.1348/0144666041501688
31. Danner, U. N., Aarts, H. & de Vries, N. K. Habit vs. intention in the prediction of future behaviour: The role of frequency, context stability and mental accessibility of past behaviour, (in eng). *Br. J. Soc. Psychol.* **47**(Pt 2), 245–65 (2008). https://doi.org/10.1348/014466607x230876
32. Roumeliotis, K. I. & Tselikas, N. D. ChatGPT and Open-AI models: A preliminary review. *Future Internet* **15** (6). https://doi.org/10.3390/fi15060192
33. Hasani, A. M. et al. Evaluating the performance of generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. *Eur. Radiol.* (2023). https://doi.org/10.1007/s00330-023-10384-x
34. Kalyan, K. S., Rajasekharan, A. & Sangeetha, S. AMMUS: A survey of transformer-based pretrained models in natural language processing. *arxiv prepreint* (2021).
35. Sorensen, T. et al. *An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels*. 819–862 (2022).
36. Dang, H., Mecke, L., Lehmann, F., Goller, S. & Buschek, D. How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. arxiv prepreint (2022).
37. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. arxiv prepreint (2022).
38. Fu, Y., Peng, H., Sabharwal, A., Clark, P. & Khot, T. Complexity-based prompting for multi-step reasoning. arxiv prepreint. https://doi.org/10.48550/arXiv.2210.00720
39. Shin, J., Jo, E., Yoon, Y. & Jung, J. A system for interviewing and collecting statements based on intent classification and named entity recognition using augmentation. *Appl. Sci.* **13**, 20, https://doi.org/10.3390/app132011545
40. Shen, Y. et al. HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face. arxiv prepreint, (2023).
41. Balcombe, L. AI Chatbots in digital mental health, *Informatics* **10** (4). https://doi.org/10.3390/informatics10040082
42. Nordgren, I. & Svensson, G. E. *Prompt Engineering and its Usability To Improve Modern Psychology Chatbots, Bachelor* (School of Electrical Engineering and Computer Science, 2023).
43. Bran, A. M. et al. ChemCrow: Augmenting large-language models with chemistry tools, arxiv prepreint (2023).
44. Park, J. S. et al. Generative agents: Interactive simulacra of human behavior. arxiv prepreint (2023).
45. Shao, Y., Li, L., Dai, J. & Qiu, X. Character-LLM: A trainable agent for Role-Playing. Arxiv Prepreint, (2023).
46. Zekun Moore, Z. P. et al. Junran Peng, RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arxiv prepreint, (2023).
47. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models, *Nature* **623** (7987), 493–498 (2023). https://doi.org/10.1038/s41586-023-06647-8
48. Schunk, D. H. Social cognitive theory. *APA Educ. Psychol. Handb.* **1**, 101–123 (2012).
49. Balikas, G. Comparative analysis of open source and commercial embedding models for question answering. In *Presented at the Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*,, Birmingham, United Kingdom. (2023). https://doi.org/10.1145/3583780.3615994
50. Guo, S. & Sanner, S. Probabilistic latent maximal marginal relevance, presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland, 2010. [Online]. https://doi.org/10.1145/1835449.1835639
51. Zhang, S. et al. Personalizing dialogue agents: I have a dog, do you have pets too? 01/22 2018.
52. White, J. et al. *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. (2023).
53. Wang, J. et al. Review of large vision models and visual prompt engineering, *Meta-Radiol.* **1** (3), 100047 (2023). https://doi.org/10.1016/j.metrad.2023.100047
54. Ray, P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.*, **3**, 121–154 (2023). https://doi.org/10.1016/j.iotcps.2023.04.003
55. Lo, J. H., Huang, H. P., Chen, Y. C. & Chen, J. H. Memory robot design: A new perspective from human brain model and large Language model. *IEEE Access.* **13**, 28539–28549. https://doi.org/10.1109/ACCESS.2025.3538889 (2025).
56. Lu, S. R., Lo, J. H., Hong, Y. T. & Huang, H. P. Implementation of engagement detection for human–robot interaction in complex environments. *Sensors* **24** (11). https://doi.org/10.3390/s24113311
57. Safdari, M. et al. *Personality Traits in Large Language Models*. (2023).
58. Johnson, J. A. Measuring Thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *J. Res. Pers.*, **51**, 78–89 (2014). https://doi.org/10.1016/j.jrp.2014.05.003
59. J. M. Bland and G. A. Douglas, Statistics notes: Cronbach's alpha, *BMJ* **314** (7080), 572 (1997). https://doi.org/10.1136/bmj.314.7080.572.
60. Bubeck, S. C. et al. Yi., sparks of artificial general intelligence: Early experiments with GPT-4. Arxiv Preprint. https://doi.org/10.48550/arXiv.2303.12712 (2023).
61. Laura Ruis, A. F., Bradley, H., Rahmani, H. A., Choe, K. W. & Grefenstette, E. Tim Rocktaschel, Do LLMsselectively encode the goal of an agent's reach? In *Proceedings of the First Workshop on Theory of Mind in Communicating Agents*, Honolulu, Hawaii, USA., (2023).
62. Le, M. B. & Nickel, Y-L. Maximilian, Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 5872–5877. (2019).
63. Natalie Shapira, M. L., Alavi, S. H., Zhou, X., Choi, Y. & Goldberg, Y. Maarten Sap, Vered Shwartz, Clever hans or neural theory of mind? Stress testing social reasoning in large language models. arxiv preprint (2023).
64. Maarten, R. L., Sap, D., Fried, Y. & Choi Neural theory-of-mind? On the limits of social intelligence in large LMs. arxiv preprint, (2023).
65. Chen, Y., Huang, G. & Wang, K. Effects of font size, stroke, and background on the legibility of Chinese characters in virtual reality for the elderly, (in eng). *Ergonomics*, pp. 1–11 (2024). https://doi.org/10.1080/00140139.2024.2392798
66. Chen, Y. & Wu, Z. A review on ergonomics evaluations of virtual reality, (in eng). *Work* **74** (3), 831–841 (2023). https://doi.org/10.3233/wor-205232

Acknowledgements

This research was funded by the National Science and Technology Council under Grant number 112-2221-E-002-149-MY3. All members from Robotics Lab who provided insight and expertise are greatly appreciated.

Author contributions

Jia-Hsun Lo and Jie-Shih Lo wrote the main manuscript text and collected the data from experiment. Jia-Hsun Lo prepared all the figures and tables. the Han-Pang Huang supervised the research and provided insights of modification. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-01528-8.

Correspondence and requests for materials should be addressed to H.-P.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.