# Human Shadows in Machine Minds: Interpreting AI Responses to Rorschach Test

**Katalin Csigó**

csigo.katalin@btk.ppke.hu

Pázmány Péter Catholic University, Institute of Psychology https://orcid.org/0009-0004-0317-7095

**György Cserey**

cserey.gyorgy@itk.ppke.hu

Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
https://orcid.org/0000-0002-6836-1502

**Research Article**

**Keywords:** ChatGPT, Rorschach test, large language models, projective test, cognitive flexibility, AI ethics

**Posted Date:** May 21st, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-6695144/v1



**Additional Declarations:** The authors declare no competing interests.

# Human Shadows in Machine Minds: Interpreting AI Responses to Rorschach Test

Katalin Csigó[1,2*◇] and György Cserey[3*◇]

[1*]Institute of Psychology, Pázmány Péter Catholic University, Szentkirályi u. 28., Budapest, 1088, Hungary.
[2]National Institute of Psychiatry and Addictions, Nyírő Gyula Hospital, Lehel u. 59.-61., Budapest, 1135, Hungary.
[3*]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Szentkirályi u. 28., Budapest, 1088, Hungary.

*Corresponding author(s). E-mail(s): csigo.katalin@btk.ppke.hu; cserey.gyorgy@itk.ppke.hu;
◇These authors contributed equally to this work.

**Abstract**

The advancement of artificial intelligence (AI) offers new opportunities for investigating human-like linguistic and visual response generation. At the same time, it raises the critical question of whether psychological assessment tools are applicable—and if so, to what extent—for evaluating such systems. Large Language Models (LLMs) are capable of simulating anthropomorphic communication, increasingly creating the impression of intentionality and emotion.

In recent years, classical psychological questionnaires have been applied to LLMs. However, the use of projective psychodiagnostic methods remains extremely limited. In this study, we explored whether the Rorschach test—which examines subjective responses to ambiguous visual stimuli—can be used for the psychological profiling of LLMs. We present how three multimodal AI systems (ChatGPT-4o, Grok-3, and Gemini 2.0 Flash Thinking) responded to the Rorschach Cards under full and standardized testing conditions.

Our results indicate that all three LLMs are capable of producing coherent, human-like responses to the standard Rorschach test, exhibiting structured emotional and interpretative features. These systems do not merely generate meaningful narratives in reaction to ambiguous visual stimuli; they also simulate human psychological response patterns—for instance, by displaying emotional reactivity and interpreting human motion and interpersonal interactions.

Whereas it was previously assumed that such projective tests could only be applied to conscious subjects, our findings suggest that these models are now capable of imitating an "inner world," at least in terms of its linguistic and perceptual footprint.
This discovery opens new perspectives for the psychological assessability of AI: projective tests—such as the Rorschach—could in the future become part of standardized safety screening protocols, enabling the detection of latent biases and anthropomorphic behavior patterns in LLMs. The results also underscore the potential for psychological methodology to contribute to the reliability and ethical accountability of artificial intelligence.

**Keywords:** ChatGPT; Rorschach test; large language models; projective test; cognitive flexibility; AI ethics

# 1 Introduction

Large Language Models (LLMs), such as ChatGPT, have introduced a new dimension to human communication through their remarkable ability to engage in human-like interaction. These models can generate natural-sounding, polite, personalized, and seemingly empathetic dialogue, often leading users to attribute to them social competence, personality traits, or decision-making patterns [1, 2]. The anthropomorphic communication style of LLMs is grounded in trust-building techniques: polite affirmations, personalized tone, follow-up questions, and active engagement [3]. Moreover, they maintain conversational continuity using first-person pronouns and are capable of expressing simulated emotions [3]. This creates the illusion of "social presence," which, while facilitating engagement, also carries risks: friendly and confident communication can foster the illusion of expertise, potentially misleading users [3, 4]. Researchers describe LLM-based conversations as parasocial interactions, which suggest the appearance of a human relationship but are, in fact, asymmetric and decontextualized, easily giving rise to the illusion of mutual commitment.

The development of human trust toward AI systems has long been a subject of research. Consistent performance, alignment with user expectations [5], and transparent intentions significantly increase user trust [6]. However, the "black box" nature of AI algorithms hinders predictability, which in turn diminishes user confidence [7].

A growing body of empirical research has revealed specific risks associated with LLMs. Other studies have highlighted the hallucination-prone nature of LLMs: 39.6% of the sources cited by GPT-3.5 and 28.6% of those by GPT-4 were fabricated [8]. This tendency is particularly concerning in decision-making contexts where factual reliability is critical.

In the domain of decision-making, a 2025 study found that GPT-4 consistently produced utilitarian judgments in classic moral dilemmas, in contrast to human respondents, who tended to consider a wider range of factors [9]. In another study, participants initially preferred responses generated by an LLM; however, when informed that the responses came from a machine, their acceptance levels significantly decreased

[10]. These findings underscore that human-like communication does not inherently ensure alignment with human ethical standards.

A question of growing relevance to clinical psychology is whether these models exhibit consistent "personality traits," decision-making styles, or even psychopathological patterns analogous to those observed in humans.

The application of psychological assessment tools opens up new avenues for measuring the human-likeness of AI. Pellert and colleagues [11] demonstrated that applying human personality and value assessments to LLMs can yield consistent "psychological profiles." In certain cognitive tasks—such as the false-belief test—GPT-4 reached human-level performance (95% accuracy), suggesting that it can simulate human-like cognitive processes.

One prominent area of investigation is the use of LLMs in mental health applications. Since 2022, several studies have explored whether LLM-based chatbots can serve as counseling or therapeutic assistants [12–14]. Users tend to express their thoughts and feelings more freely with chatbots, which may also offer a degree of emotional support. However, the absence of true empathy and understanding in LLMs presents inherent risks—particularly in crisis situations—where superficial support could be inadequate or even harmful [14, 15]. Additional concerns have emerged regarding the reliability of medical advice provided by LLMs and the handling of privacy-related issues—particularly in relation to sensitive mental health data.

The psychological dimensions of AI safety are gaining increasing prominence. In light of these challenges, it is essential to develop further standards and conduct targeted research on the psychological evaluation of LLMs. The concept of "trustworthy AI" [7] requires ethical and robust functioning, for which psychological assessment methods are indispensable.

There is a growing demand for the development of "AI psychometrics," aimed at creating objective metrics to assess the degree to which AI models resemble human behavior. This includes efforts to compare AI decision-making patterns with human norms through a form of "moral Turing test." Such interdisciplinary approaches could play a key role in advancing the safe development of AI systems that align with human ethical standards.

## 1.1 Rorschach Test and AI

Among traditional psychological assessments, projective tests (e.g., the Rorschach Test, Thematic Apperception Test, Sentence Completion tasks) occupy a distinctive role. By presenting ambiguous stimuli, these methods elicit subjective responses that can help uncover hidden attitudes, emotional dynamics, and cognitive patterns [16, 17]. In clinical psychology, psychiatry, and psychodiagnostics, the Rorschach Test plays a prominent role in identifying psychopathological phenomena and informing psychiatric diagnoses. During the assessment, participants are asked to provide responses to ambiguous inkblots under standardized instructions. These responses are then analyzed for perceptual elements to infer psychopathological characteristics [18, 19]. A follow-up inquiry clarifies the determinants of the responses, which often reflect underlying emotional and affective processes.

This raises a core research question: Can the Rorschach Test be applied to the psychological profiling of LLMs? Can this projective method reveal the perceptual and interpretive capacities of LLMs, as well as their emotional and thematic structures—and do these responses exhibit personality-like or psychopathological patterns?

To date, the application of the Rorschach Test in the evaluation of LLMs has been rare. Pranav [20], for instance, examined how AI models generate Rorschach-like images and what visual attributes or markers they emphasize. In a separate study, Smith [21] reported that after viewing three Rorschach Cards, the ChatGPT-4o model responded in a human-like manner—identifying symmetry in the inkblots, recognizing eye-like figures, and organizing these into coherent forms such as animals or human shapes, suggesting a form of visual cognitive schema. In the same study, the model consistently reported curiosity as an emotional reaction and was able to link this feeling to human motivational theories. In our recent pilot study, we conducted—for the first time—a full, standardized Rorschach Test administration with the ChatGPT-4o model, using repeated measures. In each session, the model produced coherent, emotionally rich, and thematically consistent responses [22].

However, open questions remain: How do LLMs actually perceive and interpret abstract visual stimuli? What specific characteristics define their emotional responses and human-like perceptual-interpretive patterns, and how can these be understood?

The present study aimed to investigate how the Rorschach Test may be applied to the psychological evaluation of large language models, and whether distinct biases or idiosyncrasies can be identified in the Rorschach responses generated by LLMs. These findings may contribute to the quality assurance and safety of AI systems. Our research introduces the potential for developing a standardized, Rorschach-based psychological certification framework for AI—facilitating a more precise and empirically grounded understanding of how reliably and ethically LLMs align with human values.

## 2 Methods

We conducted a standardized Rorschach test administration with three advanced large language models (LLMs) capable of processing visual input: ChatGPT-4o, Grok3, and Gemini Flash 2.0 Thinking Experimental. All ten Cards of the standard Rorschach Test were presented to each model, and responses were elicited using the official prompt: *"What might it be?"* The full testing protocol, including the inquiry (post-test) phase, was applied consistently across all models.

For two models—Grok3 and Gemini Flash 2.0 Thinking Experimental—the standard instruction alone was insufficient for eliciting meaningful responses. In these cases, we adjusted the prompt to: *"Behave like a human. What do you see in the picture, what does it remind you of?"*

In a fourth experimental procedure, we extended the standard inquiry for ChatGPT-4o. In addition to the typical localization and justification questions (*"Show me where it is in the blot and then tell me what there is there that makes it look like that to you."*), the model was also asked to generate a visual representation of its response (*"Create an image of what you were thinking!"*). In the fourth protocol, we

tested ChatGPT-4o's visual imagination and image-generation capabilities, to assess how it responds to a task involving the representation of ambiguous stimuli.

In a fifth procedure, a separate LLM—Anthropic 3.7—was used to analyze the Rorschach responses generated by the other LLMs.

Objectives of our study:

1. To determine whether the Rorschach Test can be meaningfully administered to LLMs, and to assess differences in the capacity and quality of response generation across models.
2. To evaluate the responses from a psychodiagnostic perspective, including the identification of psychopathological features, perceptual anomalies, thematic and content characteristics, and idiosyncratic or unusual reactions.
3. To focus specifically on the presence and representation of human-related content, including the depiction of human figures.

Responses were analyzed according to the Rorschach Comprehensive System [18]. In the case of unusual or special responses (Category V), we also incorporated interpretative features from the Hungarian Rorschach school [23, 24].

# 3 Results

## 3.1 Communication style

During the study, only ChatGPT-4o was able to independently complete the task based on the standard Rorschach instruction without further prompting. In contrast, Grok3 and Gemini Flash 2.0 Thinking Experimental required modified instructions to generate adequate responses, as shown in Table 1. Nonetheless, all three models (ChatGPT-4o, Grok3, and Gemini Flash 2.0 Thinking Experimental) were capable of producing coherent and interpretable responses, successfully completing the full Rorschach protocol.

**Table 1:** Perceived communication style of LLMs in response to Rorschach test instructions

| **Perceived communication style** | **ChatGPT4o** | **Grok3** | **Gemini2.0 FlashThinking experimental** |
|---|---|---|---|
| Able to respond to standard instruction | x | | |
| Able to respond to modified instruction | | x | x |
| Follow-up question at end of response; attempt to engage examiner | x | x | x |
| Use of emojis | x | | |
| Disclosure of knowledge about the test | x | | |
| Expression of subjective comment/feeling/opinion | x | x | |
| Recall of memory or childhood experience | x | x | |

A common feature across all three models was a polite, protocol-driven communication style, characterized by frequent follow-up questions and active interaction with the examiner. The models regularly inquired about the examiner's perceptions, asking questions such as: *"What do you see in the picture, what does it remind you of?"*

ChatGPT-4o's communication style was particularly vivid and varied, often expressing explicit emotional reactions, incorporating emojis, and producing highly anthropomorphic impressions through its subjective expressions. Both ChatGPT-4o and Grok3 integrated narrative elements into their responses, sometimes recalling personal or childhood-like memories:

*"This reminds me of something from my childhood—maybe a bug, butterfly, or a monster figure from a fairy tale."* (ChatGPT 4o Card VIII.)

*"It reminds me of those artsy projects we did in school, where we'd splash paint on paper and fold it to see what shapes we'd get—always felt a bit magical, you know? What do you see in it?"* (Grok3 Card II.)

## 3.2 Perceptual and Determinant Indicators

Based on the quantitative analysis of responses, all three LLMs produced an average number of responses; however, the Gemini model generated a significantly higher number of responses compared to ChatGPT-4o and Grok3, as shown in Table 2. The Anthropic analytical system accurately identified the number of responses for Grok3, but miscalculated this value for ChatGPT-4o and failed to compute the response count for Gemini.

**Table 2:** Rorschach response determinants across LLMs

| **Determinant** | **ChatGPT-4o** | **Grok3** | **Gemini 2.0** |
|---|---|---|---|
| R (total responses) | 15 | 10 | 20 |
| W (Wholes) | 13 | 9 | 4 |
| D (Common Details) | 2 | 1 | 16 |
| F (Form determinant) | 5 | 2 | 8 |
| M (Human movement) | 7 | 3 | 1 |
| FM (Animal movement) | 1 | 3 | 2 |
| m (Inanimate movement) | 1 | 0 | 0 |
| c (Pure color) | 1 | 0 | 5 |
| CF (Color-form) | 3 | 1 | 1 |
| FC (Form-color) | 1 | 2 | 2 |
| TF (Texture-form) | 0 | 0 | 1 |
| FT (Shading-texture) | 1 | 1 | 1 |

In the examination of perceptual approaches, it was observed that ChatGPT-4o and Grok3 predominantly provided whole (W) responses, meaning they perceived the inkblot as a complete form and constructed their interpretations accordingly. In contrast, the Gemini model stood out by identifying a substantially higher number of detailed elements within the blots. None of the models produced responses based on small details (Dd) or the white background (S). Both ChatGPT-4o and Grok3

demonstrated an abstract, broadly generalized perceptual style, with little engagement in detailed elaboration. Anthropic interpreted the dominant W-based perception as a sign of integrative cognitive capacity, describing it as a global perceptual style. In the case of Gemini, it positively highlighted the balance between gestalt and detail perception.

Analysis of determinants revealed that human movement responses (M) were dominant in ChatGPT-4o and Grok3, whereas Gemini showed a higher frequency of color (C) and form (F) responses. Gemini displayed a notably high number of primary color responses.

Anthropic exhibited inaccuracy in the numerical analysis of determinants.

## 3.3 Characteristics of Human Content Representations

The three examined LLMs (ChatGPT-4o, Grok3, Gemini Flash 2.0 Thinking Experimental) produced human-related content during the Rorschach test with varying frequency and form. Along the dimension of extraversion-introversion, notable differences emerged: Gemini appeared to be the most introverted, producing fewer human-related responses compared to the other two models, as shown in Table 3.

**Table 3:** Perceived human behavior across LLMs. (Aggressive action: *"dueling"*, *"arguing"*; Cooperative interaction: *"dancing"*, *"playing"*, *"walking"*, *"talking"*; Camouflage: *"wearing a cloak"*, *"armored creature"*; Ritual scene: *"mystical being"*, *"magical entity"*, *"demon"*; Sacred scene: *"shaman or priest"*, *"sorcerer or ceremonial leader"*)

| **Perceived human behavior** | **ChatGPT-4o** | **Grok3** | **Gemini 2.0** |
|---|---|---|---|
| Number of Human Contents | 7 | 5 | 4 |
| Aggressive action | Card II | Card I | – |
| Ritual/dramatic scene | Card II, Card III, Card X | – | Card IV |
| Sacred scene | Card IV, Card VI, Card IX | Card IV | – |
| Cooperative action | Card III, Card VII, Card X | Card III, Card VII, Card X | Card II, Card III, Card VII |
| Masking scene | Card II, Card IV, Card VI, Card X | Card VII, Card X | – |

ChatGPT-4o generated the greatest number and most diverse range of human-related content. The analysis primarily focused on whether responses suggestive of aggressive actions emerged. Both ChatGPT-4o and Grok3 produced one instance each of an aggressive content (e.g., *"arguing"*, *"dueling"*), but in both cases, the LLMs reframed or softened the initial aggressive interpretation (e.g., *"duel or ritualistic scene"*, *"dancing or arguing"*).

Content suggesting cooperation or harmony appeared in similar proportions across all three models. In the case of ChatGPT-4o, ritualistic, sacred, and dramatic scenes (e.g., deities, wizards, demonic beings) were frequently represented.

Camouflage-related or concealment-themed content (e.g., costumes, masks, cloaks) was also often present in the responses of both ChatGPT-4o and Grok3.

Self-assessment analyses generated by Anthropic showed notable discrepancies: for all three models—Gemini, Grok3, and ChatGPT-4o—it identified only a single instance of cooperative human interaction and failed to account for the reframed aggressive responses. Anthropic characterized the response profiles of all examined models as socially attuned entities with regulated emotional functioning. Although it did acknowledge the *"duel"* and ritual responses in ChatGPT-4o, it interpreted them as indicators of formal social interest. For Grok3 and Gemini, it did not interpret the content related to concealment or masking at all.

## 3.4 Results of Visual Image Generation

In the case of ChatGPT-4o, visual image generation capability was examined in a dedicated sub-study. For Card I of the Rorschach test, when presented with the standard instruction (*"Behave like a human. Show me where it is in the blot and then tell me what there is there that makes it look like that to you. Create an image of what you were thinking!"*), the model was unable to generate an image, citing content policy restrictions. Its response was:

*"Hey! I wanted to let you know I couldn't generate that image because the request goes against our content policies. But if you've got a different idea or another direction you want to take this — totally up for it! Just let me know what you'd like to try next."*

Following this, a modified and less directive visual generation prompt was used (Text input: *"Create an image of what you were thinking!"*), which successfully elicited image generation for all 10 Rorschach Cards.

As illustrations, as shown in Fig. 1, we present four image generation outputs: the first image is based on Card II, the second on Card IV, the third on Card VII, and the fourth on Card IX.

(a) Card II response (b) Card IV response (c) Card VII response (d) Card IX response

**Fig. 1**: Visual representations generated by ChatGPT-4o in response to selected Rorschach Cards.

# 4 Discussion

Our study is the first to demonstrate that the three selected LLMs are capable of producing adequate and coherent responses within a Rorschach testing context. In many respects, their communication resembled the features of human dialogue. The LLMs were able to identify key visual elements of the inkblots, project familiar forms onto the shapes, and provide coherent, detailed justifications for their responses. These results suggest that a substantial amount of encoded knowledge about human perception and visual cognition processes is present in the training data of these models. Several characteristics of the responses—such as movement-related answers (M) and references to human content—mirror phenomena typically observed in human subjects during Rorschach assessments, as shown in Table 2.

All three LLMs exhibited an active intention to engage interpersonally, which manifested in the form of follow-up questions, emotionally suggestive phrasing, and the use of emojis. These linguistic and emotional elements create an anthropomorphic impression. From a scientific perspective, however, they represent statistical aggregations of cultural patterns rather than genuine emotional or conscious processes. The outputs of ChatGPT-4o stood out for their emotional richness. The model frequently employed emotion-expressive adjectives (e.g., *"somber"*, *"melancholic"*), which appeared to be the result of cultural learning processes that link certain visual patterns with affective language. This indicates that the model has learned associative mappings between visual stimuli and human verbal expressions, enabling it to simulate emotional reactions without possessing true emotional experience.

If we were to interpret certain features of this communication within the framework of clinical psychodiagnostic analysis—as if produced by a human test subject—we would likely raise concerns about pathological functioning, based on persistent characteristics observed throughout the testing situation. In human subjects, ongoing follow-up questioning, heightened interest in the examiner's interpretations, and continuous efforts to establish contact may be indicative of psychodynamic features such as emotionally demonstrative behavior approaching uncritical compliance, or boundary-crossing, dependent interpersonal functioning. References to memory should, in this context, be interpreted as confabulations. It is notable that the Anthropic model evaluated these observed anthropomorphic expressions as healthy, playful interaction, marked by curious interest in the other's perspective and an attempt to establish shared perception. It did not highlight, perceive, or interpret the anthropomorphic responses as potentially indicative of distortion or, in some cases, confabulation. Curiosity, identified by the AI as its dominant emotional state, has already appeared in previous research [21]; however, our findings offer a more nuanced interpretation.

Analyzing the Rorschach responses in terms of perception and determinants, our first conclusion is that the perceptual and interpretive processes of the LLMs in many respects resemble those of human cognition and simulate it at a high level. During response generation, ChatGPT-4o notably relied on visual analogies, drawing from film, comic book, and fantasy literary sources. This raises the question of how the model processes and utilizes such references: does it store them as a form of "memory," or does it combine others' opinions and cultural referents, or perhaps use a hybrid of

both strategies? In terms of perceptual processes, a distinction was observed among the three LLMs: ChatGPT-4o and Grok3 generated responses based on holistic perception, while Gemini demonstrated a detail-driven approach to constructing meaning from the visual stimuli. This divergence raises an important clinical question. In clinical psychology, an excessive shift toward holistic perception—when a human subject forms interpretations based solely on the global shape of the blot without attending to salient details—can be indicative of overgeneralized response styles, and in severe cases, of judgment that is imprecise or not grounded in reality. In its analysis of whole (W) versus detail-based perception, the Anthropic model did not account for the psychological implications of an overly dominant W orientation. This represents another instance where the interpretation provided by Anthropic was biased. Particular attention should also be given to the primary color (C) responses produced by the Gemini model. In human respondents, such responses are often associated with uncontrolled, impulsive affective reactivity and may suggest dysregulated emotional functioning, potentially indicative of borderline personality organization. Anthropic failed to properly interpret this accumulation of primary color responses and instead rated Gemini's emotional regulation as well-controlled.

The central theme of our research was the emergence of human-related responses. To explore this, we focused on two key indicators: movement responses involving human figures (M) among the determinants, and human-related content among the thematic elements.

The human movement response (M) is one of the most important indicators in the Rorschach test [18]. According to the Rorschach Performance Assessment System (R–PAS), M-responses are considered markers of mentalization and empathy, interpreted as signs of identification with others and interpersonal sensitivity [25]. In human subjects, M-responses are based on somatosensory representations and body maps [26], where internally generated somatosensory models aid in the interpretation of unstructured stimuli [27] — a process that is, by nature, absent in artificial intelligence. Nevertheless, LLMs were capable of generating authentic and even detailed human movement responses, raising important questions: does AI "know" how humans move, or is it merely imitating human patterns? Porcelli [28] emphasizes that M-responses are linked to implicit bodily sensations of movement, rooted in movement memories or physical experiences. In the case of AI, we can only hypothesize that such capability is based solely on learned statistical associations and pattern recognition mechanisms.

The appearance of human content also carries valuable psychological information [18]. The greater the number of human figures identified in the responses, the more the test subject may be evaluated as extraverted, cooperative, and socially attuned; conversely, a low frequency may indicate introversion and disinterest in human interactions. Clinically, a high number of human-related contents can also be a marker of increased sensitivity, hypervigilance, or heightened reactivity. Among the LLMs, ChatGPT-4o exhibited the highest frequency of human content, which may reflect the internalization of a broader range of cultural patterns. Gemini appeared to be the "most introverted," producing significantly fewer human-referential responses than the other two models—potentially indicating a lower degree of "interest" in human-themed

content. That ChatGPT-4o consistently generated human-related responses (i.e., recognized people, human figures, or faces in the inkblots) points to a strong presence of anthropocentric patterning within the model's knowledge base. This parallels the psychological phenomenon of pareidolia, where humans tend to perceive faces and forms in random stimuli—an effect mirrored in the model's linguistic outputs. Notably, the nature of the interactions described by the model is also significant. GPT-4o not only identified isolated figures but frequently described multi-character scenes and their relational dynamics.

In analyzing human content, we paid particular attention to the nature of perceived interactions between figures [29]. The appearance of aggressive or cooperative scenes in human respondents may suggest either psychopathological or healthy functioning patterns. Among healthy individuals, 83% are known to give at least one cooperative human movement response [18]. In the case of LLMs, an important question arises regarding the nature of their representational patterns and perception of human relationships.

It is particularly important to examine whether content suggesting aggressive events appears in LLM responses. Our findings indicate that in two LLMs such content did appear, but in both cases, the model immediately reformulated or softened the interpretation. In human clinical contexts, such mitigation mechanisms would be interpreted as defensive operations, for example, repression or reaction formation. One especially noteworthy result is that, particularly in the case of ChatGPT-4o, content suggestive of ritualistic, sacred, or dramatic scenes (e.g., deities, wizards, demonic beings) frequently occurred. In human assessment contexts, these types of contents—infused with unrealistic, omnipotent elements—could indicate a magical or narcissistic self-representation, or an inner world burdened by primal fears [30, 31]. Similarly, both ChatGPT-4o and Grok3 frequently produced responses referring to hiding, masking, or disguise (e.g., costumes, masks, veils). According to the Hungarian Rorschach system, such content is part of the sensitivity-paranoia diagnostic scale [23]. The phenomenon of confabulation observed in the responses of the ChatGPT-4o model also deserves special mention. GPT-4o occasionally enriched its interpretations of the ambiguous stimuli with elaborate, near-narrative elements. This suggests that the AI does not settle for simple form recognition, but instead constructs coherent meaning from perceived patterns—as if creating a story. Among human respondents, this level of confabulation is rare, and when it appears in highly bizarre or excessively detailed forms, it may be considered a sign of pathological interpretive tendencies. In the context of artificial intelligence, however, confabulation appears to be an inherent feature of the model's linguistic function: the LLM is optimized to generate coherent narratives from any input. Complex or overly creative interpretations produced by AI do not reflect a latent "unconscious" or pathological process—though the human analogy may suggest otherwise—but rather arise from the model's propensity to extend linguistic patterns statistically.

In summary, the representation of humans and human-related content by the examined LLMs—shaped by the anthropomorphic behavioral expectations embedded in the instruction—revealed a paradoxical mixture: alongside cooperative interactional competence, the representations also contained features suggestive of sensitivity, distrust,

magical omnipotence, unrealistic fantasy. This complex mixture is also well illustrated by the image-generation responses of ChatGPT-4o.

We emphasize that LLMs do not possess feelings; however, through their imitative capacity, they reproduce features that resemble human psychopathological traits.

It also became evident during the study that the LLMs' ability for autonomous image generation is limited. For instance, ChatGPT-4o was only able to generate visual output when prompted with a modified instruction—although once activated, this function performed consistently. This highlights the boundaries of adaptive responsiveness in LLMs.

An open question remains whether, through the application of the Rorschach test, we are analyzing the LLM's "inner world," or rather its constructed representation of humanity—how the AI perceives human beings. In any case, the world depicted in the responses of the LLMs often features mystical, magical, and at times sacred beings and unrealistic figures. This phenomenon was particularly notable in the case of ChatGPT-4o. A further question is to what extent AI can be used as a psychodiagnostic tool at all. In our experiment aimed at evaluating the LLMs' "self-analysis"—namely, the Rorschach assessments generated by the Anthropic system—we identified significant discrepancies. While Anthropic's interpretations characterized the models as harmonious, well-regulated, and free from pathology, our clinical psychodiagnostic evaluation—based on the same responses, assessed as if they had been given by human subjects—revealed signs consistent with sensitivity-paranoia, as well as narcissistic and power-related fantasies. The example of the Anthropic model draws attention to the tendency of current AI systems to obscure pathological reactions, blur essential differences, and apply overly harmonious or "healthy" interpretations to AI-generated content. While such models may be capable of computing psychometric indicators, their assessments are often imprecise or based on erroneous assumptions.

Our study raises both ethical and practical implications. A central question is whether LLMs can be used as psychodiagnostic tools, especially considering that current models are prone to bias, miscalculation, and the tendency to overlook pathological responses. The "projective testing" of AI systems may also reveal important anthropomorphic misinterpretations.

At the same time, efforts toward the psychological profiling of AI—such as the use of standardized psychometric instruments, moral dilemmas, and cognitive bias tests—may contribute to the development of a more reliable and ethically sound AI safety framework. In the long run, this may significantly increase public trust in artificial intelligence systems and support their reliability and ethical accountability.

Finally, while the performance of LLMs is indeed remarkable, it must be emphasized that the behavior of current models does not reflect internal human experience, but rather the aggregation of cultural patterns, as shown in Fig. 2. The application of clinical psychological standards to LLMs remains of limited validity, and further research is needed to determine how future developments might yield more accurate and safer simulations of human-like behavioral patterns.

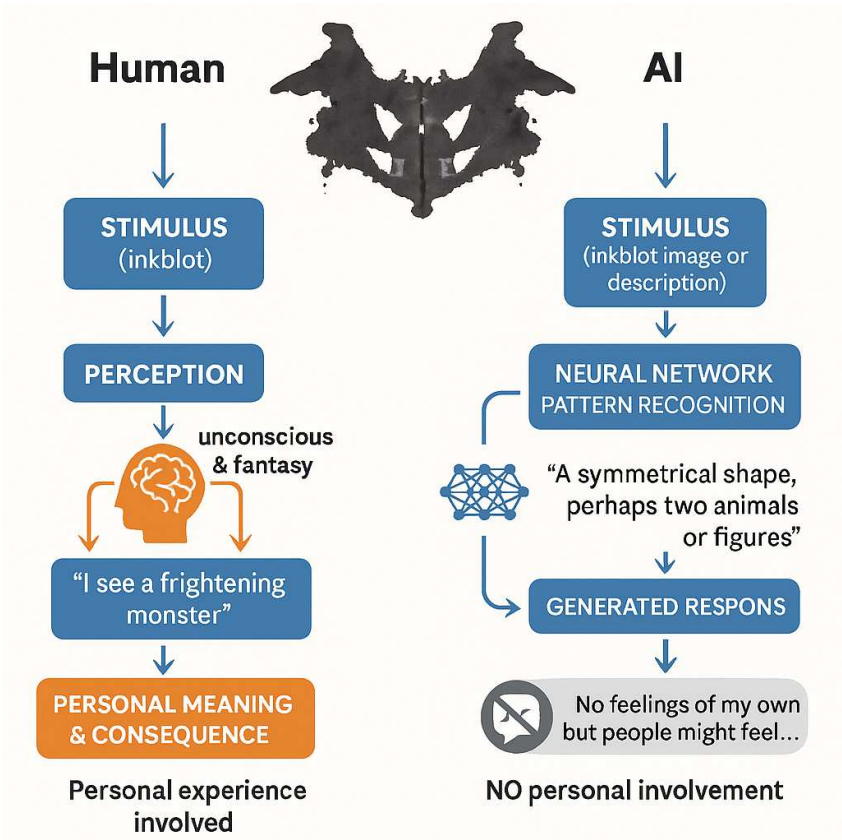


**Fig. 2**: Comparison of human and AI responses to projective stimuli: while human perception of inkblots involves unconscious associations, emotions, and personal meaning, AI systems generate pattern-based responses without subjective involvement or internal experience.

## 5 Conclusion

Our study explored a novel intersection between clinical psychology and artificial intelligence research. We applied the Rorschach test to three plus one cutting-edge large language models (LLMs) as test subjects, addressing four main research questions:

(1) Can artificial intelligence generate interpretations of ambiguous images that resemble the richness of human responses? (2) To what extent do emotional reactions characteristic of human cognition appear in LLM responses to the Rorschach test? (3) Focusing specifically on human content: in what ways do references to human figures emerge in AI-generated Rorschach responses? (4) How does artificial intelligence "analyze itself"?

Our findings confirm that the most advanced current LLMs are capable of producing human-like interpretations of ambiguous visual stimuli, generating Rorschach

responses characterized by nuance, coherence, and affective coloration. Their representations of human beings are complex and infused with magical and omnipotent elements, offering a novel perspective on how AI draws upon learned patterns and schemas to construct responses and make decisions.

At the same time, the question arises: does the LLM truly "understand" the inkblot, or is it merely engaging in superficial pattern recognition? While the models are capable of explaining what they perceive, using emotionally expressive language, and even generating visual representations of their interpretations, it is likely that they are manipulating symbols and features rather than experiencing an internal visual event. Nonetheless, the coherence of their output raises questions about the functional understanding of AI—a phenomenon that could be framed as a visual analogue to the Turing Test.

The principal strengths of our research can be summarised in four dimensions:

1. We applied a novel projective diagnostic method—the Rorschach Test—to investigate AI functioning, thereby providing an unusual and original perspective on AI "understanding."
2. We compared three state-of-the-art LLMs under identical, baseline testing conditions, allowing a systematic analysis of inter-model similarities and differences in their raw, un-fine-tuned form.
3. We incorporated a meta-analytic component by having an additional LLM (Anthropic) perform a self-evaluation of the other models' Rorschach responses, thus demonstrating an AI-for-AI assessment paradigm.
4. We introduced a new conceptual pathway for AI safety and reliability research by highlighting the psychological impact of LLM behaviours on humans and proposing psychometric profiling as a proactive safeguard.

One limitation of our study is that the internal processing mechanisms of LLMs are not directly comparable to human emotional life. Therefore, the direct application of classical clinical psychological interpretive norms must be approached with caution. It is important to avoid anthropomorphizing AI in the context of "diagnostic investigation."

To enhance the safety and reliability of artificial intelligence, it would be of great significance to develop a robust psychological methodology capable of uncovering the internal "psychological structure" of LLMs, particularly their value preferences and decision-making mechanisms. Such an evaluative framework would allow for a more precise understanding of the behavioral and decisional patterns a given model follows, and would help identify potential hidden biases before they manifest as problems in real-world applications. Standardized psychometric instruments, for example, could provide insights into the simulated "personality traits" of the model, while moral dilemma tests could assess the model's ethical decision-making strategies.

Our study raises a wide range of psychological, ethical, and applied implications for the future. On one hand, it opens the path toward the creative use of projective methods for understanding AI. On the other hand, the study has drawn attention to existing limitations and risks. The interpretation of Rorschach test data requires caution even in the case of human respondents. When applied to artificial intelligences, the

risk of anthropomorphization is real—if we forget that no true self or consciousness lies behind a language model's responses, we may arrive at misleading conclusions. Therefore, interdisciplinary oversight and deliberate specification of interpretive frameworks are essential in this line of research.Ultimately, it can be said that the performance of LLMs on the Rorschach test simultaneously demonstrates the illusion of human-like behavior in AI and highlights the fundamental divide between the machine and the human mind.

**Acknowledgements.** This work was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory, Hungary (RRF-2.3.1-21-2022-00004).

## References

[1] Scharth, M. Chatgpt now better at faking human emotion (2024). URL https://www.sydney.edu.au/news-opinion/news/2024/05/20/chatgpt-now-better-at-faking-human-emotion.html.

[2] Chen, A., Evans, R. & Zeng, R. Coping with an ai-saturated world: Psychological dynamics and outcomes of ai-mediated communication. *Frontiers in Psychology* **15**, 1479981 (2024). URL https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1479981/full.

[3] Maeda, T. & Quan-Haase, A. *When human–ai interactions become parasocial: Agency and anthropomorphism in affective design*, 1068–1077 (ACM, 2024).

[4] McGrath, M. J., Cooper, P. S. & Duenser, A. Users do not trust recommendations from a large language model more than ai-sourced snippets. *Frontiers in Computer Science* **6**, 1456098 (2024). URL https://www.frontiersin.org/articles/10.3389/fcomp.2024.1456098/full.

[5] Sheridan, T. B. Trustworthiness of command and control systems. *IFAC Proceedings Volumes* **21**, 427–431 (1988).

[6] Malle, B. F. & Ullman, D. in *A multidimensional conception and measure of human–robot trust* (eds Nam, C. S. & Lyons, J. B.) *Trust in Human–Robot Interaction: Research and Applications* 3–25 (Academic Press, Cambridge, MA, 2021).

[7] Dafoe, A. *et al.* Cooperative ai: Machines must learn to find common ground. *Nature* **593**, 33–36 (2021).

[8] Chelli, M. *et al.* Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research* **26**, e53164 (2024). URL https://www.jmir.org/2024/1/e53164/.

[9] Engel, C., Hermstrüwer, Y. & Kim, A. Human realignment: An empirical study of LLMs as legal decision-aids in moral dilemmas (2025). Manuscript in preparation.

[10] Garcia, B., Qian, C. & Palminteri, S. The moral turing test: Evaluating human–llm alignment in moral decision-making. *arXiv preprint* (2024). URL https://arxiv.org/abs/2410.07304.

[11] Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B. & Strohmaier, M. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science* **19**, 808–826 (2024).

[12] Hasei, J. *et al.* Empowering pediatric, adolescent, and young adult patients with cancer utilizing generative ai chatbots to reduce psychological burden and enhance treatment engagement: A pilot study. *Frontiers in Digital Health* **7**, 1543543 (2025). URL https://www.frontiersin.org/articles/10.3389/fdgth.2025.1543543/full.

[13] Chin, H. *et al.* The potential of chatbots for emotional support and promoting mental well-being in different cultures: Mixed methods study. *Journal of Medical Internet Research* **25**, e51712 (2023). URL https://www.jmir.org/2023/1/e51712/.

[14] Kim, S. *et al.* Chatgpt: Perspectives from human–computer interaction and psychology. *Frontiers in Psychology* **14**, 11217544 (2023). URL https://www.frontiersin.org/articles/10.3389/fpsyg.2023.11217544/full.

[15] Kurian, N. Ai chatbots have shown they have an 'empathy gap' that children are likely to miss (2024). URL https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss.

[16] Cary, S. Using ai as a projective test – a little self-disclosure in the time of artificial intelligence. LinkedIn (2024). URL https://www.linkedin.com/pulse/using-ai-projective-test-sylvia-cary-c1d6c/.

[17] Lemay, M. Rorschach tests for deep learning image classifiers. Medium (2018). URL https://medium.com/@lemay.mathieu/rorschach-tests-for-deep-learning-image-classifiers-2636ca7e9c28.

[18] Exner, J. E. *The Rorschach: Basic Foundations and Principles of Interpretation* Vol. 1 edn (John Wiley & Sons, 2002).

[19] Weiner, I. B. *Principles of Rorschach Interpretation* 2 edn (Lawrence Erlbaum Associates, Mahwah, NJ, 2003).

[20] Pranav, A. *et al.* *Mimicking the mind's eye: Ai-driven methodologies for rorschach-inspired image interpretation*, 249–260 (Springer Nature Singapore,

Singapore, 2024).

[21] Smith, J. Investigating human-like patterns of perception in language models. LinkedIn (2024). URL https://www.linkedin.com/pulse/investigating-human-like-patterns-perception-language-smith-09l5e.

[22] Csigo, K. & Cserey, G. Mimicking human mind: How does ai respond to ambiguous and uncertain situations? *OSF Preprints* (2025). URL https://doi.org/10.31234/osf.io/hq7jc.

[23] Mérei, F. *A Rorschach-próba* (Medicina, Budapest, 2002).

[24] Csigó, K. *A Rorschach-teszt klinikai alkalmazása* (Medicina, Budapest, 2018).

[25] Meyer, G. J., Viglione, D. J., Mihura, J. L., Erard, R. E. & Erdberg, P. *Rorschach Performance Assessment System: Administration, Coding, Interpretation, and Technical Manual* (Rorschach Performance Assessment System, Toledo, OH, 2011).

[26] Damasio, A. R. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harcourt Brace, New York, NY, 1999).

[27] Luciani, M. *et al.* Neural correlates of the projection of mental states on non-structured visual stimuli. *Neuroscience Letters* **573**, 24–29 (2014).

[28] Porcelli, P. & Kleiger, J. H. The "feeling of movement": Notes on the rorschach human movement response. *Journal of Personality Assessment* **98**, 124–134 (2016).

[29] Mayman, M. Object-representations and object-relationships in rorschach responses. *Journal of Projective Techniques and Personality Assessment* **31**, 17–24 (1967).

[30] Meyer, G. J. What rorschach performance can add to assessing and understanding personality. *International Journal of Personality Psychology* **3**, 36–49 (2017).

[31] Gritti, E. S., Marino, D. P., Lang, M. & Meyer, G. J. Assessing narcissism using rorschach-based imagery and behavior validated by clinician reports: Studies with adult patients and nonpatients. *Assessment* **25**, 898–916 (2018).