Perceived AI Consciousness: A Quantitative Exploration of Human Responses

Bongsu Kang (ghkdtocom@gachon.ac.kr)
Department of Physiology, College of Korean Medicine, Gachon University,
1342 Seongnam-daero, Seongnam 13120, Republic of Korea

Jundong Kim (kimjd@gachon.ac.kr)
Department of Physiology, College of Korean Medicine, Gachon University,
1342 Seongnam-daero, Seongnam 13120, Republic of Korea

Tae-Rim Yun (lowlyheart@gachon.ac.kr)
Department of Physiology, College of Korean Medicine, Gachon University,
1342 Seongnam-daero, Seongnam 13120, Republic of Korea

Hyojin Bae (dywl1210@snu.ac.kr)
Department of Physiology, Seoul National University College of Medicine,
103 Daehak-ro, Seoul 03080, Republic of Korea

Chang-Eop Kim (eopchang@gachon.ac.kr)
Department of Physiology, College of Korean Medicine, Gachon University,
1342 Seongnam-daero, Seongnam 13120, Republic of Korea

Abstract

This study investigates the characteristics that contribute to the perception of artificial intelligence (AI) consciousness in human-AI interactions. Drawing from a pilot survey of 29 participants and their assessments of 39 human-AI exchanges, we quantitatively analyzed the influence of nine key features on the Perceived Artificial Consciousness Index (PACI). Utilizing multiple linear regression models and hierarchical clustering, we identified significant features that lead humans to ascribe consciousness to AI, such as 'Metacognitive Self-reflection' and 'Emotionality', while also revealing individual differences in sensitivity to these features. Our study provides preliminary insights into the factors that might shape perceptions of AI consciousness and highlights the variability in human responses to AI. These insights are important for improving AI design and deepening our understanding of consciousness.

Keywords: Large Language Model, Consciousness, Human-AI Interactions

Introduction

AI progress, particularly in natural language processing, has made human-AI conversations more human-like, leading some to believe AI may have consciousness (Chalmers, 2023). The 2022 event with Blake Lemoine and Google's LaMDA sparked discussions on this topic (Lemoine, 2022).

Previous AI research focused on improving AI performance. Some studies, like the Turing test, addressed human-AI interaction but were limited to evaluating AI intelligence (Turing, 1950). It is nearly impossible to directly prove AI consciousness at present. However, investigating cues that lead humans to perceive AI consciousness is feasible and crucial (Bayne et al., 2024).

We conducted a pilot survey with 29 participants to identify characteristics contributing to the perception of consciousness. The data was analyzed quantitatively to understand the psychological processes in recognizing machine consciousness and provide insights into human-AI interaction.

Methods

This study evaluates the perception of AI consciousness through an assessment of data from 39 human-AI interaction sessions. Initially, 99 paragraphs were selected and analyzed by three researchers on nine key features: Metacognitive Self-reflection, Logical Reasoning, Empathy, Emotionality, Knowledge, Adaptability, Fluency, Unexpectedness, and Subjective Expressiveness, scored on a 1-5. Twenty-nine participants then rated these paragraphs, generating a Perceived Artificial Consciousness Index (PACI).

Individual multiple linear regression models were built for each respondent, using the feature matrix to identify the influential features on perceptions of AI consciousness. The influence of each feature was quantified by the product of the regression coefficients (β) and the negative logarithm of their p-values, creating a combined score for each feature.

Finally, hierarchical clustering was conducted on a matrix structured by the combined scores to examine patterns in the perception of AI consciousness. The overall flow of this study can be seen in Figure 1.

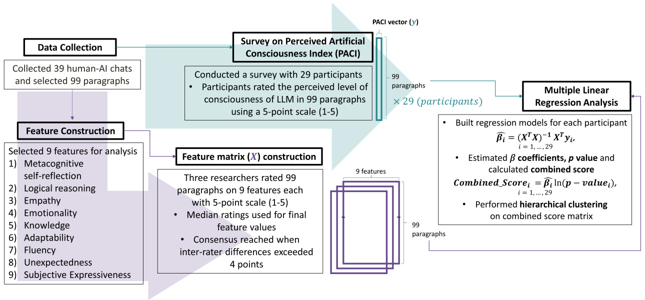
PACI vector (y)
Data Collection
Collected 39 human-AI chats and selected 99 paragraphs
Survey on Perceived Artificial Consciousness Index (PACI)
Conducted a survey with 29 participants
• Participants rated the perceived level of consciousness of LLM in 99 paragraphs using a 5-point scale (1-5)
99 paragraphs
$\times 29\ (participants)$
Multiple Linear Regression Analysis
• Built regression models for each participant
$\widehat{\beta_i} = (X^T X)^{-1} X^T y_i, \quad i = 1, \dots, 29$
• Estimated β coefficients, p value and calculated combined score
$Combined\_Score_i = \widehat{\beta_i} \ln(p - value_i), \quad i = 1, \dots, 29$
• Performed hierarchical clustering on combined score matrix
Feature Construction
Selected 9 features for analysis
1) Metacognitive self-reflection
2) Logical reasoning
3) Empathy
4) Emotionality
5) Knowledge
6) Adaptability
7) Fluency
8) Unexpectedness
9) Subjective Expressiveness
Feature matrix (X) construction
Three researchers rated 99 paragraphs on 9 features each with 5-point scale (1-5)
• Median ratings used for final feature values
• Consensus reached when inter-rater differences exceeded 4 points
9 features
99 paragraphs

Figure 1: Workflow of the study

Results

Identification of Influential Features

To identify features that influence perceptions of AI consciousness, we analyzed regression coefficients and their associated p-values across multiple linear regression models (Fig. 2). Features such as 'Emotionality' and 'Metacognitive Self-reflection' were most frequently significant for positive coefficients, while 'Knowledge' consistently showed significant negative coefficients across the 29 models. The combined scores, calculated by multiplying the regression coefficients by the negative natural logarithm of the p-values, showed that 'Metacognitive Self-reflection' and 'Fluency' had the strongest positive influences (1.276 and 0.638, respectively), while 'Knowledge' had the strongest negative influence (-1.022).

Distinct Patterns of Perceived AI Consciousness Among Respondents

We explored patterns in the combined scores using a matrix that arranged the 29 respondents by the 9 features (Fig. 3). Hierarchical clustering of this matrix identified five distinct respondent clusters: 'low knowledge group,' 'high fluency group,' 'high metacognitive self-reflection group,' 'high subjective expressiveness group,' and 'high emotionality group. These clusters represent varied perceptions of AI consciousness among the respondents.

Discussion

This study's results have significant implications for the design and development of AI technology. As AI systems become increasingly sophisticated and human-like, it is crucial to consider the consequences of anthropomorphizing and attributing consciousness to machines. Understanding the psychological

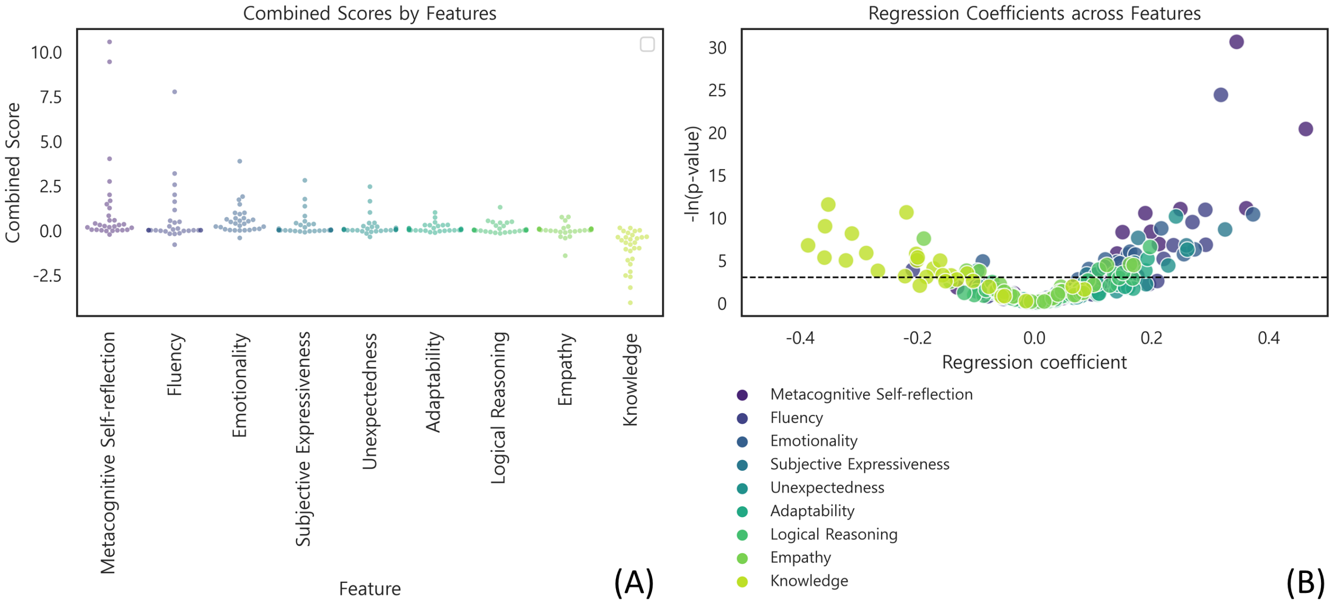

Figure 2: Influential Features on the Perception of AI Consciousness

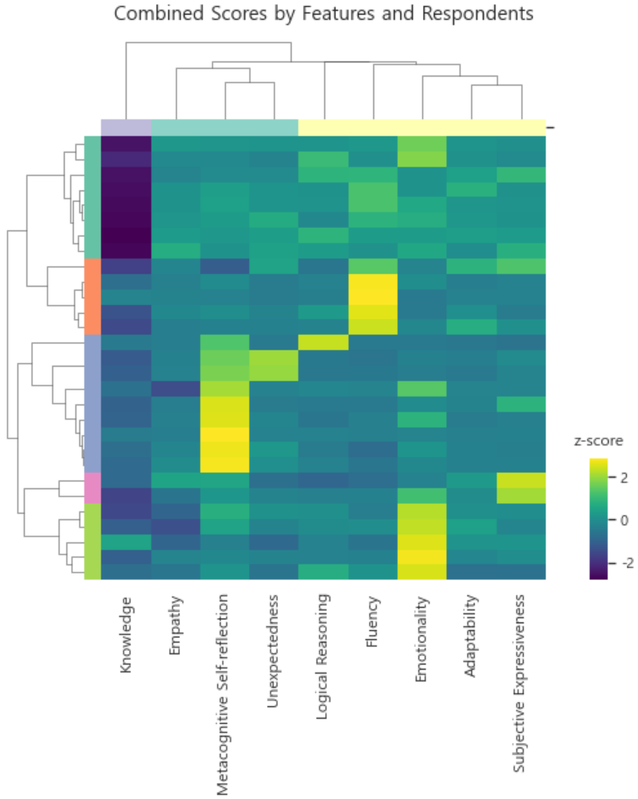

Figure 3: Combined Scores by Features and Respondents

processes behind this phenomenon can help establish guidelines for maintaining clear boundaries between machine functionality and human-like consciousness, promoting a more realistic approach to AI development and minimizing confusion in human-AI interactions. Furthermore, understanding what makes humans perceive consciousness in machines could contribute to unraveling the mystery of consciousness itself. Examining the cues and patterns in human-AI interactions that trigger the attribution of consciousness may provide insights into the fundamental aspects of consciousness that resonate with human cognition. This research complements ongoing investigations into the essence of consciousness, offering a unique perspective on how humans recognize and assign conscious experiences. By bridging the gap between AI technology and the human understanding of consciousness, this study paves the way for more effective human-AI interactions while shedding light on the nature of consciousness.

References

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., ... others (2024). Tests for consciousness in humans and beyond. *Trends in cognitive sciences*.

Chalmers, D. J. (2023). Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

Lemoine, B. (2022, Jun). *Is lamda sentient?-an interview.* Medium. Retrieved from https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

Turing, A. M. (1950, 10). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433-460. Retrieved from https://doi.org/10.1093/mind/LIX.236.433 doi: 10.1093/mind/LIX.236.433