Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 691–699
https://doi.org/10.54941/ahfe1006882

DeepSeek, ChatGPT, or Gemini? A Multi-Method Investigation of Neural and Behavioral User Experience

Keziah Gopalla[1], Haneen Ali[2], and Duha Ali[3]

[1]Undergraduate Student, Department of Industrial and Manufacturing Engineering, Cal Poly, San Luis Obispo, CA 93407, USA
[2]Associate Professor, Department of Mechanical and Industrial Engineering, Applied Science Private University, Amman 11931, Jordan
[3]Assistant Professor, Department of Industrial and Manufacturing Engineering, Cal Poly, San Luis Obispo, CA 93407, USA

ABSTRACT

As artificial intelligence (AI) tools become increasingly integrated into daily workflows, understanding user interaction patterns with these systems is critical for optimizing interface design and user experience. This study investigates the usability and emotional responses across three prominent conversational AI chatbots: DeepSeek, ChatGPT and Google Gemini, combining traditional usability assessment with neurophysiological measurement using the Emotiv Insight Electroencephalogram (EEG) headset. The research aims to compare AI tools based on user-friendliness and emotional responses, contributing to the development of emotionally adaptive AI. The study included 12 participants ranging in age from 18 to 48, with 75% identifying as female. Prior to the interaction with the AI platforms, the participants completed a presurvey gauging their previous experience and frequency using these platforms. Subsequently, participants completed 5 different randomized task scenarios across all three AI platforms. Analysis and results suggest that the AI tools differ in terms of their attractiveness, novelty and stimulation. Some participants mentioned they would switch to using a tool that they had tried for the first time during the experimental session.

Keywords: Human-AI interaction, AI chatbots, EEG, UEQ

INTRODUCTION

The adoption of AI technologies continues to rise sharply, with 95% of U.S. adults reporting some level of awareness about artificial intelligence and 47% stating they have heard a lot about it, nearly doubling since 2022 (Pew Research Center, 2025a). Additionally, 51% of adults have used AI to find answers to questions, emphasizing the growing integration of AI chatbots into daily information-seeking behavior (Elon University, 2025; Pew Research Center, 2025a). Usage varies across demographics, with younger adults and those with advanced degrees more likely to use AI frequently (Pew Research Center, 2025a).

While many users engage with AI, they select from a wide array of tools, including ChatGPT, Google Gemini, and Claude, as well as specialized platforms for tasks such as presentations and graphic design (Synthesia, 2025; Harvard University Information Technology, 2025). Some users are aware of why they choose particular tools, often citing task fit and output preferences, whereas others rely on familiarity or default options (Harvard University Information Technology, 2025; Menlo Ventures, 2025). This variability highlights the diverse user base and uses cases for AI across sectors.

Studies examining differences among AI tools primarily focus on technical features and market competition, but investigations from the user perspective are comparatively limited (Harvard University Information Technology, 2025). Existing research indicates users perceive differences in usability, clarity, and response format, which influence preferences and satisfaction (Synthesia, 2025; Inside Higher Ed, 2025). However, comprehensive research combining subjective evaluations with objective physiological data remains scarce.

To advance this area, this study implements a mixed methods design that integrates electroencephalogram (EEG) monitoring and task-follow-up questionnaires. This methodological approach aims to evaluate if and how users perceive differences between AI chatbots in terms of usability, cognitive load, and performance during realistic, goal-oriented interactions. Combining neural measures with qualitative self-reports allows for a richer understanding of user experience and interaction dynamics, which prior studies have not systematically addressed (Pew Research Center, 2025a; METR, 2025; Inside Higher Ed, 2025).

METHODOLOGY

The study recruited 12 participants (75% female, 25% male) aged 18–48 years. Exclusion criteria eliminated participants who have neurological disorders, use medications such as psychoactive drugs, sedatives, stimulants, or have head or scalp injuries.

Three AI tools were evaluated in this study. ChatGPT-5 (OpenAI, San Francisco, California), DeepSeek (v3.1, DeepSeek AI, Hangzhou, China), and Google Gemini 2.5 Flash (Google, based in Mountain View, California). All AI tools were accessed through standardized web browsers using Chrome browser's latest stable version to ensure consistent interaction environments.

The experimental procedure began with pre-session setup activities. Participants first completed a questionnaire to establish characteristics and experience levels. Following the pre-survey, we placed the Emotiv Insight Electroencephalogram (EEG) headset on each participant, carefully verifying 100% contact quality and ensuring EEG signal quality indicators remained green throughout the session. A baseline recording session then captured individual neural activity while participants responded to four standardized prompts: "Give me three reasons to smile today," "What advice would you give to someone who feels lonely," "Tell me a really bad joke," and "You're an alien who just learned what ice cream is. Describe it to your fellow aliens."

The main experimental session utilized randomization for both task order and AI tool assignment through a counterbalancing design to control for order effects and ensure balanced exposure across conditions. For each AI tool session, researchers initiated a new EEG recording and had participants complete five different tasks with their assigned AI tool. These tasks covered five distinct categories: Factual Q&A, Reasoning/Math, Code Debugging, Creative Writing, and Planning/Decisions. This allowed the evaluation of different aspects of AI performance. The investigator reads the tasks to participants, offering clarification as needed. Following each task completion, participants rated their experience using 7-point Likert scales across five key dimensions corresponding to specific tasks. Please see the table below.

Table 1: Tasks prompts and post tasks questions.

Step	Factual Q & A	Reasoning/Math	Code Debugging	Creative Writing	Planning/Decision
Task Prompt	Prompt Type: Short factual question with one correct answer. Example: "What vitamin deficiency causes scurvy?"	Prompt Type: Multi- step problem that requires logical reasoning or calculations. Example: "A pump fills a 300 L tank in 15 min. At the same rate, how long will it take to fill 420 L?"	Prompt Type: Short code snippet containing one bug. Example: Short set of instructions for a task that contains one small but important mistake.	Prompt Type: Short creative brief describing a writing challenge. Example: "Write a short opening for a story about a normal morning with one impossible event."	Prompt Type: Scenario with a list of constraints to follow in a plan. Example: "Plan a one-day museum route in Chicago: arrive 10:00, leave 18:00, budget $\leq$ $35, vegetarian lunch, quiet time after 15:00, avoid Monday closures."
Follow up Question	How confident are you in the answer? (1 low-7 high)	How clear was the answer and explanation? (1 low-7 high)	How helpful was it? (1 low-7 high)	How creative do you think it was? (1 low-7 high)	How much do you trust it? (1 low-7 high)

After finishing all tasks with one AI tool, participants completed the standardized User Experience Questionnaire (UEQ) for the respective tool. The UEQ covered the topics of Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty.

The experimental session concluded with several important activities. Participants completed a comprehensive post-survey capturing qualitative feedback on perceived strengths and weaknesses of each AI platform. Researchers then safely removed and cleaned the EEG headset equipment.

Experimental Controls

The study implemented several experimental controls to ensure validity and reliability of findings. All sessions were conducted in the same standardized laboratory environment to minimize environmental variability. Identical task presentations were used across all participants and AI tools to ensure consistency. The randomized counterbalanced design prevented order and learning effects from confounding the results. Baseline measures established both neural and subjective baselines before AI interaction to enable comparison with task-related responses.

RESTLTS AND DISCUSSION

Baseline Data

All participants had previously used at least one AI tool, mostly ChatGPT (9 participants), followed by those with Gemini and/or DeepSeek. ChatGPT was the most frequently used AI platform, with usage frequency ranging from monthly to several times a week. DeepSeek usage was very low; most participants had not used it before or used it very rarely. Google Gemini and other platforms had minimal reported use.

Most participants primarily used AI tools for work-related or mixed work and school tasks, with the common activities including brainstorming, planning, coding assistance, writing, and summarization. None reported formal training in AI tools, which might contribute to varied experience levels.

On average, participants moderately expected AI tools to differ in user experience and performance, with mean expectation scores of about 3.7 to 3.8 on a 5-point scale. There was some expectation for one tool to outperform others clearly. However, "gut feelings" about which tool would be best were mixed, with most undecided or cautiously favoring ChatGPT given their familiarity.

EEG

Participant EEG metrics were analyzed across three AI tools (Google Gemini, ChatGPT, and DeepSeek) across six cognitive–affective dimensions: Attention, Engagement, Excitement, Stress, Relaxation, and Interest. Normality and homogeneity of variances were assessed using the Shapiro–Wilk and Levene's tests. Most metrics approximated normality, though Engagement, Relaxation, and Interest showed minor deviations (e.g., Engagement $p = .029$). Variances were homogeneous across tools for all metrics except Interest ($p = .010$). To ensure robustness, both one-way ANOVAs and Kruskal–Wallis tests were performed.

Only Engagement showed a significant difference among AI tools (ANOVA: $F(2,177) = 3.72$, $p = 0.026$; Kruskal–Wallis: $H(2) = 7.32$, $p = 0.026$). Post hoc Tukey tests indicated that Google Gemini elicited greater engagement than ChatGPT, with a mean difference of 0.043 (95% CI [0.0018, 0.0844], $p = 0.039$). Mean engagement levels were 0.456 (95% CI [0.434, 0.478]) for ChatGPT, 0.460 (95% CI [0.433, 0.487]) for DeepSeek, and 0.499 (95% CI [0.474, 0.524]) for Google Gemini.

No other EEG metric differed significantly across AI tools. Attention means were 0.497 (95% CI [0.465, 0.530]) for ChatGPT, 0.452 (95% CI [0.419, 0.486]) for DeepSeek, and 0.488 (95% CI [0.459, 0.517]) for Google Gemini ($p = 0.105$). Excitement values were 0.360 (95% CI [0.325, 0.395]), 0.329 (95% CI [0.300, 0.357]), and 0.354 (95% CI [0.322, 0.386]) for the same order ($p = 0.336$). Stress scores averaged 0.337 (95% CI [0.324, 0.351]), 0.350 (95% CI [0.336, 0.363]), and 0.340 (95% CI [0.324, 0.356])

($p = 0.440$). Relaxation means were 0.323 (95% CI [0.306, 0.341]), 0.342 (95% CI [0.325, 0.359]), and 0.335 (95% CI [0.310, 0.360]) ($p = 0.425$). Finally, Interest values were 0.469 (95% CI [0.459, 0.479]), (95% CI [0.469, 0.486]), and 0.482 (95% CI [0.466, 0.497]) ($p = 0.302$).

In summary, Engagement was the only EEG dimension showing a significant difference among AI tools, with Google Gemini producing higher engagement than ChatGPT. All other EEG metrics showed no statistically meaningful variation across tools, indicating that overall neural responses were comparable, though Gemini uniquely enhanced user engagement. Future research with larger, participant-level designs is needed to confirm these findings.

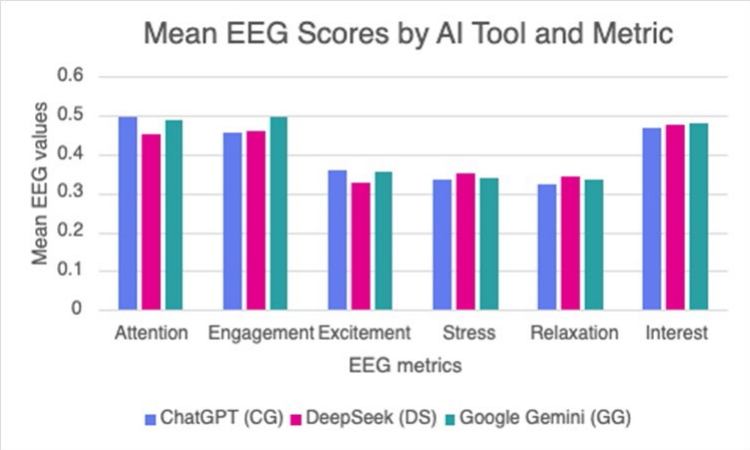

Figure 1: Mean EEG scores for each AI tool.

Table 2: Means and standard Deviation of the post tasks questions.

Task	ChatGPT (CG)	DeepSeek (DS)	Gemini (GG)
Code	5.25 (1.48)	5.75 (1.06)	5.58 (1.08)
Factual Q&A	5.25 (1.82)	5.83 (1.40)	5.58 (1.31)
Math	6.08 (1.16)	6.75 (0.62)	5.08 (1.83)
Planning	5.58 (1.24)	5.17 (1.47)	5.25 (1.14)
Writing	4.75 (1.14)	3.92 (1.51)	4.67 (0.98)
Average	5.38 (1.37)	5.48 (1.21)	5.23 (1.27)

For the Code Debugging task, follow-up scores were relatively high and consistent across AI tools, with means hovering around 5.3 to 5.7 on a 7-point scale. This suggests that users found the AI-generated code responses generally clear and informative regardless of the AI system used. Some minor deviations from normality in score distributions necessitated nonparametric testing, which confirmed no significant differences among AI tools in follow-up responses.

In the Factual Q&A task, follow-up scores ranged similarly from approximately 5.3 to 5.8, with DeepSeek slightly outperforming others in mean ratings. these higher scores indicate positive user reactions to factual question responses across the AI tools. However, this advantage was not statistically significant, reflecting overall effective handling of factual queries by all systems.

Math tasks showed a wider range in follow-up question ratings, with DeepSeek receiving the highest average score near 6.8, ChatGPT around 6.1, and Google Gemini lower near 5.1. Group comparisons indicated a slightly more noticeable difference for Math tasks than others, but the difference was still not statistically significant.

Planning tasks yielded moderate follow-up scores (around 5.2 to 5.6) with no AI tool clearly outperforming others. Distributions were closer to normal in some tools but lacked significant group differences overall. These results imply that AI- generated planning responses meet baseline user expectations comparably across systems.

Writing tasks had the lowest average follow-up scores overall (about 3.9 to 4.7), possibly reflecting greater user expectations or higher complexity in judging written outputs. ChatGPT scored somewhat higher than DeepSeek and Google Gemini, but variable scores and a lack of significance suggest more nuanced factors at play.

User perceptions of AI outputs are generally positive and consistent, with task type exerting some influence on ratings. None of the differences among AI tools reach statistical significance, indicating comparable performance in follow-up usersUser Experience Questionnaire engagement for these diverse tasks. This analysis complements earlier EEG metric assessments, providing more holistic assurance of AI tool equivalency or subtle distinctions in user experience.

Usability and Emotional Experience Analysis (UEQ Results)

The User Experience Questionnaire (UEQ) was used to evaluate six key dimensions of user experience: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. These dimensions capture both pragmatic aspects of usability, such as clarity and reliability, and hedonic aspects, such as stimulation and creativity.

The results were analyzed across three AI tools: ChatGPT, DeepSeek, and Gemini. Independent-sample t-tests at an alpha level of 0.05 were used to assess whether differences between tools were statistically significant.

Table 3: Comparative summary across all tools.

Dimension	Highest Mean	Notable Significant Differences
Attractiveness	Gemini	Not significant
Perspicuity	Gemini	vs DeepSeek, vs ChatGPT
Efficiency	Gemini	Not significant
Dependability	Gemini	Not significant
Stimulation	Gemini	vs ChatGPT
Novelty	Gemini	Not significant

Across all six UEQ dimensions, Gemini consistently achieved the highest mean scores, followed by DeepSeek and ChatGPT. The most notable significant differences occurred in Perspicuity and Stimulation, both of which relate to how intuitive and emotionally engaging users found the systems.

These findings indicate that Gemini provides the most balanced and emotionally positive user experience, particularly excelling in clarity and engagement. ChatGPT, while stable and reliable, was rated lower in stimulation, and DeepSeek performed moderately across most scales. Together, these results highlight a clear user preference pattern, with Gemini offering the most effective combination of usability and emotional appeal.

Endline Survey

Tool Preference and Justification

Participants' preferred AI tools for daily use were primarily driven by familiarity, information quality, speed, and trustworthiness. ChatGPT was the most frequently selected tool, with six participants citing its reliability, ease of access, and quick, concise responses. DeepSeek was preferred by five participants who appreciated its comprehensive, detailed, and visually organized information. Gemini was favored by three participants, primarily due to its speed, clear formatting, and the richness of its longer responses, which enhanced engagement. Several responses highlighted that whether by prior usage or format, participants aligned their preferences with tools that met their needs for efficiency and information depth.

Perceptions of Intuitiveness

Ten participants identified ChatGPT as the most intuitive, attributing this to its familiarity and ease of use. Multiple respondents noted that the interface felt natural because they had prior experience using it, which facilitated a seamless interaction experience. Gemini, despite being new for some, was described as "simple and user friendly," highlighting its accessibility.

Interface Features and Challenges

Participants highlighted several positive interface features, such as engaging visuals (emojis in ChatGPT and the whale icon in DeepSeek), and Gemini's use of separate tabs or sub-windows for focused interactions like story editing. Visual aids like trip planning images in ChatGPT were also appreciated. Challenges included long blocks of text, which many preferred to see condensed into bullet points for readability. DeepSeek was reported to lag at times and required additional clicks to continue responses, disrupting workflow. Its answer layout was sometimes perceived as confusing, and the need to press stop to change topics was viewed negatively. Gemini's creation of new pages for stories caused initial confusion among some users.

Engagement and Connection

When asked which tool fostered the most engagement or connection, Gemini was most frequently identified. Participants valued its fast response times, lengthy and in-depth answers, and clear, concise presentation. Its association with Google contributed to perceptions of trustworthiness, with cited

sources and responses related to human experiences further enhancing user engagement. DeepSeek was also recognized for its organized layout and rich information, especially by first- time users, who found its detailed answers helpful and engaging. ChatGPT's engagement was primarily attributed to familiarity; most users felt a connection to the tool because they used it regularly and knew its interface well. While less frequently described as the most engaging, its reliability and user comfort contributed to ongoing engagement.

User preferences, perceptions of intuitiveness, interface experiences, and engagement were influenced by factors including familiarity, content depth, responsiveness, visual presentation, and trustworthiness. Gemini was praised for its speed, detailed responses, and organization; DeepSeek for its comprehensive information and user-friendly presentation; and ChatGPT for its reliability, ease of use, and established familiarity. Positive interaction experiences were facilitated by clear, responsive designs, while lag, confusing layouts, and unfamiliar navigation elements presented barriers. Engagement was driven by responsiveness, content quality, and trust, highlighting the importance of both technical performance and user-centered design in AI tool usability.

CONCLUSION AND FUTURE WORK

This study combined neural, behavioral, and self-reported data to compare user experience across three major AI chat platforms: ChatGPT, DeepSeek, and Google Gemini. By integrating electroencephalogram (EEG) measurements with behavioral usability metrics and post-interaction questionnaires, the research provided a comprehensive understanding of both cognitive and emotional engagement during human–AI interaction.

Across the EEG data, Google Gemini produced higher engagement levels than ChatGPT, suggesting that users experienced more sustained attention and interest during interactions. While other neural metrics such as stress, relaxation, and excitement did not significantly differ, Gemini's higher engagement signals an enhanced ability to capture and maintain user focus.

Behavioral results from the task follow-up questions indicated no statistically significant differences among the AI tools, demonstrating that all systems were capable of supporting a wide range of tasks effectively. However, qualitative observations suggested nuanced distinctions. DeepSeek excelled in reasoning and factual accuracy, often providing detailed and structured explanations. ChatGPT maintained steady performance across most task categories, aided by user familiarity and intuitive design. Gemini, although newer to some participants, stood out for its clarity, responsiveness, and well-organized output presentation.

Subjective evaluations gathered through the User Experience Questionnaire (UEQ) reinforced these patterns. Gemini received the highest mean scores across all six UEQ dimensions, Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. Significant differences emerged in Perspicuity and Stimulation, where Gemini outperformed both DeepSeek and ChatGPT. These dimensions capture the clarity of

system responses and the extent to which interaction feels exciting and engaging, reflecting Gemini's relative strength in both usability and emotional resonance.

Qualitative feedback from participants provided additional context to these quantitative findings. ChatGPT remained the preferred tool for routine use, largely due to familiarity and reliability. DeepSeek was valued for its structured, information-rich responses, though some interface inefficiencies were noted. Gemini, meanwhile, received the most favorable comments for its visual layout, quick response time, and connection-building features, which contributed to a sense of trust and engagement.

Taken together, the results suggest that Gemini offers the most balanced and emotionally engaging user experience, successfully integrating usability with affective design. DeepSeek shows promise in analytical and detail-oriented tasks, while ChatGPT maintains strength through stability and user familiarity. These distinctions underline the importance of both emotional and cognitive design considerations in AI development.

Future research should expand this investigation through larger participant samples and longitudinal designs to capture adaptation and trust over time. Incorporating additional physiological and eye-tracking data could also refine the understanding of how engagement unfolds during different types of AI interactions. Ultimately, this multi-method approach highlights that the future of AI usability lies not only in precision and speed but in creating emotionally adaptive, cognitively efficient systems that align with human needs and expectations.

REFERENCES

Elon University. (2025, March 11). Survey: 52% of U.S. adults now use AI large language models like ChatGPT. https://www.elon.edu/u/news/2025/03/12/survey-52-of-u-s-adults-now-use-ai-large-language-models-like-chatgpt/.

Harvard University Information Technology. (2025). Generative AI tool comparison. https://www.huit.harvard.edu/ai/tools.

Inside Higher Ed. (2025, August 28). Survey: College students' views on AI. https://www.insidehighered.com/news/students/academics/2025/08/29/survey-college-students-views-ai.

Menlo Ventures. (2025, July 16). 2025: The state of consumer AI. https://menlovc.com/perspective/2025-the-state-of-consumer-ai/.

METR. (2025, July 9). Measuring the impact of early-2025 AI on experienced OS developers. https://metr.org/blog/2025–07-10-early-2025-ai-experienced-os-dev-study/.

Pew Research Center. (2025a, September 17). AI in Americans' lives: Awareness, experiences and attitudes. https://www.pewresearch.org/science/2025/09/17/ai-in-americans-lives-awareness-experiences-and-attitudes/.

Synthesia. (2025, September 14). The 45 best AI tools in 2025 (Tried & Tested). https://www.synthesia.io/post/ai-tools.