*Human Factors in Design, Engineering, and Computing*, Vol. 238, 2026, 1469–1477
https://doi.org/10.54941/ahfe1008375

# Consistency of AI-Driven Usability Evaluations for Software: Learning Management Systems as an Example

**Victor K. Y. Chan**

Macao Polytechnic University, Macao, China

## ABSTRACT

The evaluation of usability is a critical, yet resource-intensive, component of Human Systems Integration (HSI) for complex software systems like Learning Management Systems (LMSs). This study investigates the potential of using Large Language Model (LLM)-based Artificial Intelligence (AI) systems to conduct automated usability evaluations by analyzing textual user feedback. The author compared the consistency of usability assessments for 100 LMSs generated by three leading AI systems (Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash) against each other and against a baseline of human user aggregate scores. The AI systems scored seven usability attributes based on ISO/IEC 25010, 9241 Standards and usability models proposed in the literature so far: effectiveness, efficiency, error protection, accessibility, operability, learnability, and satisfaction. Principal Component Analysis (PCA) was used to derive a single usability factor score for each AI system's evaluations. Subsequent correlation analyses (Pearson, Spearman, Kendall's tau-b) revealed strong, statistically significant positive correlations between the AI-generated factor scores and the human benchmark scores, with correlation coefficients ranging from 0.459 to 0.626. Notably, the inter-AI correlations were exceptionally high, with values as strong as 0.850 between Claude-Opus-5 and Gemini-3.5-Flash. The findings suggest that LLMs possess a significant capability to perform human-like usability evaluations (or at least humans' aggregate product rating) from text, supporting their potential as a powerful tool for rapid, scalable, and cost-effective usability analysis in HSI and systems engineering. However, the moderate strength of the correlations with human benchmark scores also indicates that current AI systems should be seen as complementary to, rather than a complete replacement for, traditional human-centered evaluation methods.

**Keywords:** Artificial intelligence, Usability evaluation, Learning management systems, Human systems integration, Principal component analysis, Text analytics

## INTRODUCTION

Human Systems Integration (HSI) is a cornerstone of modern systems engineering, aiming to optimize total system performance by holistically considering human capabilities and limitations throughout the system lifecycle (Booher, 2003; Folds et al., 2008). A key activity within HSI is usability evaluation, which assesses the ease with which users can interact with a system to achieve their goals. Traditional usability evaluation methods,

Received September 25, 2026; Revised October 31, 2026; Accepted November 5, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

such as user testing, surveys, and expert inspections, are effective but often face significant challenges in cost, time, and scalability. These challenges are particularly acute for widely used software systems that evolve rapidly, like Learning Management Systems (LMSs) (Al-Maani and Bani Salameh, 2017), which are central to modern education and corporate training.

The advent of sophisticated Artificial Intelligence (AI), particularly Large Language Models (LLMs), presents a novel opportunity to address these challenges. LLMs possess advanced natural language processing capabilities, allowing them to analyze vast quantities of unstructured text—such as online user reviews and comments—to extract sentiments and evaluations on specific attributes (Singh Chauhan, 2022). This capability suggests a potential paradigm shift: using AI to perform a form of "text-analytics" usability evaluation, effectively synthesizing the collective voice of users to generate quantitative usability scores. This approach could offer a rapid, scalable, and cost-effective alternative or supplement to traditional methods.

Previous research has explored using AI for various aspects of usability analysis, such as automated and remote usability and user experience (UX) evaluation (Liu, 2025) and sentiment analysis of user feedback (Guzman and Maalej, 2014). However, the specific ability of contemporary LLMs to consistently and accurately replicate a multi-dimensional usability evaluation model on a large scale has not been extensively studied. This research aims to address this gap by investigating the following primary question: To what extent can AI systems, by analyzing textual user feedback, consistently evaluate the usability of LMSs in a manner that aligns with human benchmark assessments?

To achieve this, the author designed a study that leverages the usability model proposed by Al-Maani and Bani Salameh (2017), which defines LMS usability through seven key attributes based on ISO/IEC 25010, 9241 Standards and usability models proposed in the literature so far: (1) effectiveness, (2) efficiency, (3) error protection, (4) accessibility, (5) operability, (6) learnability, and (7) satisfaction. The analytical capabilities of three prominent AI systems, Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash were utilized, tasking them with scoring a sample of 100 LMSs on each of these seven attributes based on their analysis of global user reviews and comments. The consistency of these AI evaluations was then assessed both among the AI systems themselves and against an external benchmark of human user aggregate scores.

The study's contribution is threefold. First, it provides an empirical assessment of the capability of state-of-the-art LLMs to perform complex, multi-faceted usability evaluations. Second, it analyzes the degree of agreement and variance between different AI systems on the same evaluative task. Third, it compares AI-derived usability scores against benchmark human aggregate scores, providing a critical measure of their ecological validity. The findings are expected to inform practitioners and researchers in HSI, human factors, and systems engineering about the potential utility and limitations of AI-driven usability evaluation tools.

## METHODOLOGY

### Data and Materials

The dataset for this study was constructed by compiling the 100 top-rated LMSs listed on the software review platform Capterra as of August 24, 2026, when sorted by the "Most Reviews" criterion. This method was chosen to sample a pool of 100 LMSs with significant market presence and a substantial volume of user feedback, ensuring a rich textual dataset for AI analysis and a robust benchmark of human scores. The list included widely known platforms such as Canvas, Moodle, Google Classroom, and Docebo, among others (Capterra, 2026).

The usability model employed for evaluation was the seven-attribute framework proposed by Al-Maani and Bani Salameh (2017). This model provides a structured approach to LMS usability, encapsulating core dimensions of user interaction, including effectiveness (accuracy and completeness of user tasks), efficiency (speed of task completion), error protection (prevention and recovery from errors), accessibility (usability for users with diverse abilities), operability (ease of control and command execution), learnability (ease of learning the LMS), and satisfaction (user comfort and positive attitudes toward the LMS).

The primary evaluators were three LLMs: Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash. The prompt submitted to each of these AI systems was designed to standardize the evaluation task:

> "The following article proposes a model to evaluate the usability of Learning Management Systems (LMSs) based on seven usability attributes: (1) effectiveness, (2) efficiency, (3) error protection, (4) accessibility, (5) operability, (6) learnability, and (7) satisfaction: ... For each of these seven attributes, please give a score to each of the following 100 LMSs on a scale of 1 (worst) to 10 (best). By means of text analytics, please derive your scores from global users' textual comments on these seven attributes of the 100 LMSs as appear all around the web. Please present your scores in the form of a table..."

This prompt ensured that each AI system performed its evaluations based on a standardized conceptual framework and the same underlying source of information (global user comments), making their outputs as directly comparable as possible.

The benchmark for human user assessment was the aggregate "Overall Score" for each LMS provided by Capterra. It is important to note that the exact methodology Capterra uses to calculate this overall score is proprietary, the score sounds more like users' impressive/intuitive assessment than systemic scoring, and its direct relationship to the seven specific usability attributes from Al-Maani and Bani Salameh (2017) may not be perfect. However, it serves as a valid, real-world indicator of overall user satisfaction and perceived quality, providing a meaningful external reference for evaluating the AI's performance.

It turned out that Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash provided full evaluative scoring data to only 97 of the 100 LMSs originally shortlisted, so the analysis was confined to a dataset of 97 LMSs as such.

### Analysis

The statistical analysis was performed using SPSS. The process involved two main steps: dimensionality reduction and correlation analysis.

First, to capture the overall usability assessment of each AI system into a single, comparable metric, a Principal Component Analysis (PCA) was performed on the seven attribute scores provided by each AI system separately. This technique reduced the seven potentially correlated variables into a single, uncorrelated composite variable—the "usability factor score"—that explains the maximum variance in the original seven attribute scores. For each AI system, the first principal component, representing the underlying concept of "overall usability," was extracted to back this "usability factor score." For the sake of easy referencing, ClFactorScore, GPFactorScore, and GeFactorScore refer to these usability factor scores derived from Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash, respectively, in this paper.

Second, to assess the consistency of the AI evaluations, various correlation coefficients were calculated:

**Inter-AI Consistency**: Pearson's correlation coefficient was computed between the AI factor scores pairwise (ClFactorScore vs GPFactorScore, ClFactorScore vs GeFactorScore, and GPFactorScore vs GeFactorScore) to determine the degree of linear agreement between the different AI systems.

**AI-to-Human Consistency**: Both parametric (Pearson's) and non-parametric (Spearman's and Kendall's tau-b) correlation coefficients were calculated between each AI's factor score (ClFactorScore, GPFactorScore, and GeFactorScore) and the Capterra Overall Score. Non-parametric measures were included to account for potential non-linearity in the relationship or violations of normality.

All correlation coefficients were tested for statistical significance at the $\alpha = 0.05$ significance level, and a 95% confidence interval (CI) was calculated for each coefficient using a bias-corrected and accelerated (BCa) bootstrap method (with 2000 samples) to provide robust estimates of the effect sizes.

## RESULTS

### Descriptive Statistics and Data Screening

The analysis confirmed that all three AI systems (Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash) successfully generated full datasets of usability scores for the seven attributes across 97 out of the 100 LMSs originally shortlisted.

### Principal Component Analysis

The application of Principal Component Analysis (PCA) to the attribute scores from each AI system yielded a single, dominant component for each, explaining a substantial portion of the variance. The first principal component

of Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash respectively explained 77.3%, 73.4%, and 84.3% of the total variance.

The high loadings of all seven attributes on this single component confirmed that a single "usability factor" could effectively summarize each AI's assessment. The KMO measure of sampling adequacy (Claude: 0.908, GPT: 0.867, Gemini: 0.921) and Bartlett's test of sphericity (all $p < 0.001$) indicated the data were highly suitable for factor analysis. The component matrices showing the factor loadings are presented in Table 1.

**Table 1**: Component matrix for principal component analysis of usability scores ($N = 97$).

| Usability Attribute | Claude-Opus-5 | GPT-5.6-Sol | Gemini-3.5-Flash |
|---|---|---|---|
| | | Component 1 | |
| Effectiveness | 0.650 | 0.661 | 0.891 |
| Efficiency | 0.945 | 0.928 | 0.955 |
| Error protection | 0.933 | 0.881 | 0.907 |
| Accessibility | 0.789 | 0.812 | 0.823 |
| Operability | 0.966 | 0.917 | 0.965 |
| Learnability | 0.892 | 0.852 | 0.915 |
| Satisfaction | 0.934 | 0.917 | 0.965 |

Extraction method: Principal Component Analysis: 1 component extracted.

## Correlation Analysis: Inter-AI Consistency

The Pearson's correlation coefficients between the three AI system's usability factor scores revealed exceptionally strong, positive, and statistically significant relationships (all $p < 0.001$), as shown in Table 2.

**Table 2**: Pearson's correlations between AI system's usability factor scores ($N = 97$).

| Usability Factor Score | Usability Factor Score | | |
|---|---|---|---|
| | Claude-Opus-5 | GPT-5.6-Sol | Gemini-3.5-Flash |
| Claude-Opus-5 | 1 | 0.808*** | 0.850*** |
| GPT-5.6-Sol | 0.808*** | 1 | 0.767*** |
| Gemini-3.5-Flash | 0.850*** | 0.767*** | 1 |

*** $p < 0.001$ (2-tailed)

## Correlation Analysis: AI-to-Human Consistency

All three AI systems demonstrated a positive and statistically significant correlation with the benchmark human aggregate scores from Capterra. The results, summarized in Table 3, were consistent across parametric (Pearson's) and non-parametric (Spearman's and Kendall's tau-b) measures.

These results indicate that the AI-generated usability factor scores share a moderate-to-strong positive relationship with the benchmark human aggregate scores, supporting the validity of the AI evaluations at least in

taking the role of humans' aggregate product rating. The GPT-5.6-Sol system showed the strongest correlation (Pearson's correlation = 0.626) with the human benchmark in this sample.

**Table 3:** Correlation between AI's usability factor scores and benchmark human aggregate scores (Capterra's) ($N = 97$)

| Factor score from the AI System | Correlation with human benchmark score | Bootstrap 95% confidence interval (CI) (bias-corrected and accelerated or BCa) | |
|---|---|---|---|
| | | Lower | Upper |
| Pearson's | | | |
| Claude-Opus-5 | 0.527** | 0.376 | 0.636 |
| GPT-5.6-Sol | 0.626** | 0.499 | 0.727 |
| Gemini-3.5-Flash | 0.459** | 0.312 | 0.573 |
| Spearman's | | | |
| Claude-Opus-5 | 0.501** | 0.336 | 0.628 |
| GPT-5.6-Sol | 0.584** | 0.440 | 0.698 |
| Gemini-3.5-Flash | 0.411** | 0.231 | 0.552 |
| Kendall's tau-b | | | |
| Claude-Opus-5 | 0.359** | 0.234 | 0.474 |
| GPT-5.6-Sol | 0.434** | 0.318 | 0.543 |
| Gemini-3.5-Flash | 0.283** | 0.155 | 0.396 |

** $p < 0.01$ (2-tailed)

## DISCUSSION AND CONCLUSION

This study investigated the potential of Large Language Models (LLMs) to perform automated usability evaluations by analyzing textual user feedback for LMSs. The results provide compelling evidence for the viability of this approach. All three AI systems evaluated—Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash—demonstrated significant positive correlations with both each other and, more importantly, with an external benchmark of human user aggregate scores.

The exceptionally strong inter-AI consistency (with Pearson's correlations ranging from 0.767 to 0.850) suggest a high degree of convergence in how these advanced AI systems process and synthesize user feedback into a usability assessment. This consistency is remarkably promising, as it implies that the systems are not generating random outputs but are instead identifying a stable, underlying signal related to usability from the textual data. This finding aligns with the broader literature demonstrating the utility of LLMs for sentiment analysis and opinion mining (Singh Chauhan, 2022). The high variance explained by the first principal component (73.4% to 84.3%) for all three AI systems suggests that their evaluations of the seven usability attributes are highly cohesive, effectively reflecting a single, overarching construct of "usability." This strongly supports the theoretical framework of Al-Maani and Bani Salameh (2017) by demonstrating that

their seven attributes, as operationalized by the AI systems, indeed form a unified concept.

The relationship between the AI-derived usability factor scores and the benchmark human aggregate scores (i.e., Capterra's "Overall Score"; Pearson's correlations = 0.459 to 0.626) is particularly encouraging. It establishes a baseline of ecological validity, indicating that the AI's evaluative outputs correlate moderately well with real-world user sentiment (at least human users' aggregate product rating if not user sentiment as gauged scientifically). This suggests that AI systems can potentially serve as a rapid, cost-effective proxy for certain aspects of human evaluation. For instance, they could be used to conduct initial screenings of user feedback for large LMSs, identify potential usability hotspots, or track trends over time, thereby augmenting the capabilities of HSI professionals. The potential to "listen" to the collective voice of users at scale is a powerful tool in the systems engineering and HSI toolkit.

However, the correlations, while robust, are not perfectly strong (< 0.63). This is a crucial limitation. It suggests that current AI systems, despite their advanced capabilities, do not perfectly replicate human evaluation. This could be due to several factors. First, the AIs are synthesizing information from a vast corpus of text, which may contain noise, conflicting opinions, and exaggerated sentiments. Human users, on the other hand, might be influenced by a wider range of factors beyond the seven usability attributes, such as brand loyalty, price, or integration with other tools. Second, the Capterra's "Overall Score" is an aggregate measure whose exact composition is unknown; it may not perfectly map onto the specific seven-attribute model that the AIs were instructed to follow. Third, there may be inherent limitations of contextual misinterpretation in the AI's ability to fully understand context, sarcasm, or subtle nuances in user comments that a human evaluator would readily grasp (Liu, 2025). The wider confidence intervals for Gemini-3.5-Flash's correlations (particularly the Spearman's correlation CI of 0.231 to 0.552) suggest that its performance may be somewhat less stable compared to the other AI systems in this context.

Despite these limitations, the findings have clear implications for future research and practice. The moderate-to-strong correlations indicate that a "human-in-the-loop" approach might be the most effective path forward. In this model, AI systems could perform a first-pass analysis, identifying trends and flagging potential issues in user feedback. HSI experts could then validate these AI findings with smaller-scale, focused user studies and expert inspections, combining the scalability and efficiency of AI with the nuanced understanding of human practitioners (Folds et al., 2008). Furthermore, these results open avenues for future research. This includes investigating whether fine-tuning these LLMs on usability-specific datasets could enhance their performance, as suggested by recent work in specialized AI domains (Ravikumar and Aarthi, 2026). Additionally, future studies could aim to collect and analyze human user scores that are explicitly and directly mapped to the seven usability attributes, creating a more precise benchmark for comparison. An additional point is that in many real-life scenarios, for

example, during software selection or market analysis, what matters is not the exact numerical score but the ranking of the software systems in question. As such, the author did conduct a rank-based analysis of AI consistency in shortlisting top LMSs by usability based on basically the same dataset of the current study (Chan, 2026). This exemplifies the potential for furthering this study.

In conclusion, this study demonstrates that LLMs like Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash can perform text-analytics-based usability evaluations that are significantly consistent with human user assessments (at least humans' aggregate product rating). The remarkably high inter-AI consistency further bolsters the reliability of this approach. While not a perfect replacement for traditional HSI methods, this capability offers a powerful new tool for rapid and scalable usability analysis. As AI technology continues to evolve, its role in supporting and augmenting human factors and systems engineering processes is likely to become increasingly crucial, helping to ensure that future complex software systems are designed with the human user truly at the center.

## ACKNOWLEDGMENT

The author would like to acknowledge the use of the AI systems Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash for scoring LMSs' usability and SPSS for statistical analysis. Thanks are also due to the AI system DeepSeek-V4-Pro-0813 for its assistance in the presentation and language phrasing of this article.

## REFERENCES

Al-Maani, D. I. and Bani Salameh, H. (2017). "A generic model for evaluating the usability of learning management systems", proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing (ICC '17), Cambridge, United Kingdom. https://doi.org/10.1145/3018896.3018921

Booher, H. R. ed. (2003). Handbook of human systems integration. Hoboken, NJ: Wiley. 10.1002/0471721174

Capterra. (August 24, 2026). Best learning management system (LMS) software. Capterra Website: https://www.capterra.com/learning-management-system-software/

Chan, V. K. Y. (2026). A rank-based analysis of AI consistency in shortlisting top software systems by usability: Learning management systems as an example, Applied human factors and ergonomics (AHFE) international Volume 33 (in press).

Folds, D., Gardner, D. L. and Deal, S. (2008). Building up to the human systems integration demonstration, INSIGHT Volume 11 No. 2. 10.1002/inst.200811215

Guzman, E. and Maalej, W. (2014). How do users like this feature? A fine grained sentiment analysis of app reviews, proceedings of 2014 IEEE 22nd International Requirements Engineering Conference (RE), Karlskrona, Sweden. https://doi.org/10.1109/RE.2014.6912257

Liu, J. (2025). AI in automated and remote UX evaluation: A systematic review (2014–2024), Advances in human-computer interaction 7442179. https://doi.org/10.1155/ahci/7442179

Ravikumar, R. N. and Aarthi, S. (2026). "Fine-tuning large language models for domain-specific business applications," in: Leveraging LLMs for business innovation: practical solutions and future trends, Ismail, S. M. A. (Ed.) pp. 107–136. 10.4018/979-8-3373-3912-2.ch004

Singh Chauhan, R. (2022). Sentiment analysis of customer reviews using deep learning techniques. National College of Iceland. https://norma.ncirl.ie/6573/1/rishabhsinghchauhan.pdf