Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1428–1436
https://doi.org/10.54941/ahfe1008371

A Rank-Based Analysis of AI Consistency in Shortlisting Top Software Systems by Usability: Learning Management Systems as an Example

Victor K. Y. Chan

Macao Polytechnic University, Macao, China

ABSTRACT

This study serves as a furtherance of a prior work, which demonstrated a moderate-to-strong correlational consistency between usability factor scores generated by Large Language Model (LLM)-based Artificial Intelligence (AI) systems and a human benchmark score from Capterra. While the prior work focused on parametric and non-parametric correlations, this paper investigates the consistency of the rankings produced by three AI systems (Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash) when shortlisting top-performing Learning Management Systems (LMSs) by usability as gauged by 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. Rank-Biased Overlap (RBO) was employed to compare the AI rankings against each other and against Capterra's human benchmark score. The results show that all AI systems exhibit moderate-to-high ranking agreement with each other (RBO up to 0.878). However, the agreement between the AI rankings and the human benchmark score is substantially lower, diminishing from a moderate level at a shallow depth for top 10 LMSs (RBO ~ 0.7) to a rather weak level for the full list of 97 LMSs (RBO < 0.2). These findings indicate that while AI systems are highly consistent among themselves in ranking LMSs by their usability, their rankings somewhat diverge from the human benchmark as more LMSs are considered. This suggests that AI may be a viable tool for rapidly identifying a shortlist of potential top-usability LMSs but is not yet a reliable replacement for comprehensive human evaluation. Such findings on rankings offer more practicality than the prior work on correlations.

Keywords: Rank-biased overlap, Artificial intelligence, Usability evaluation, Learning management systems, Ranking consistency

INTRODUCTION

In a prior work (Chan, 2026), the author established that Large Language Models (LLMs) like Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash can generate usability scores from textual user feedback on Learning Management Systems (LMSs) that correlate moderately well with a benchmark of human user aggregate scores. The Pearson correlation coefficients between the artificial intelligence-(AI-)generated factor scores and Capterra's human benchmark scores ranged from 0.459 to 0.626, indicating a significant,

Received September 25, 2026; Revised October 31, 2026; Accepted November 8, 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/

positive linear relationship. This suggested that AI systems possess a notable capacity to capture the general sentiment and perceived usability of LMSs from textual user feedback.

However, in many real-world applications, such as software system selection or market analysis, the primary interest lies not in the exact numerical score, but in the *ranking* of options—specifically, which LMSs are the top contenders by usability. For instance, a Human Systems Integration (HSI) (Booher, 2003) specialist or a corporate procurement team may need to shortlist a few high-potential LMSs for further, in-depth evaluation. The ability of an AI to correctly identify and rank these top candidates is a different, and arguably more practical, measure of its utility than correlational consistency over the entire dataset (Webber et al., 2010).

Rank consistency is a distinct construct from score correlation. Two evaluators could have a strong positive correlation in their scores but produce notably different rankings, especially when scores are clustered together. Therefore, a more nuanced analysis is required to determine if AI evaluations can effectively serve as a screening tool. This is the central focus of this article.

Rank-Biased Overlap (RBO) is an established metric designed specifically for this purpose. Unlike set-based metrics like Jaccard similarity, RBO accounts for the order of items in a ranking. Furthermore, it includes a persistence parameter p that allows the researcher to control the weight given to the top of the list versus the bottom, making it ideal for evaluating the consistency of "top-k" recommendations (Webber et al., 2010). RBO has been widely used in information retrieval and recommendation systems to compare search engine results, user preferences, and recommender algorithm outputs (Chapelle and Zhang, 2009; Yilmaz et al., 2008).

This paper extends the analysis from the prior work by applying RBO to the usability factor scores derived from the three AI systems and the Capterra human benchmark scores. The primary objectives are to:

1. Assess the ranking agreement between the three AI systems pairwise.
2. Assess the ranking agreement between each AI system and the Capterra human benchmark.
3. Analyze how this ranking agreement changes as different depths of the ranked list (top 10, top 20, top 50, and the full list of 97) are considered.
4. Evaluate the impact of the persistence parameter p on the observed agreement.

By addressing these objectives, the author aims to provide a more comprehensive understanding of the practical utility of AI for automated usability evaluation and shortlisting.

METHODOLOGY

Data and Materials

This study utilizes the same dataset as the aforesaid prior work. The dataset 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 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, based on ISO/IEC 25010, 9241 Standards and usability models proposed in the literature so far, defines LMS usability through attributes: (1) effectiveness, (2) efficiency, (3) error protection, (4) accessibility, (5) operability, (6) learnability, and (7) satisfaction.

The primary evaluators were three LLM AI systems: Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash. The prompt submitted to each AI system was designed to standardize the evaluation task. The AI systems were tasked with scoring each of the 100 LMSs on a scale of 1 to 10 for each of the seven usability attributes, based on their analysis of global user's textual comments. The following is an excerpt from the prompt:

> "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..."

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, and its direct relationship to the seven specific usability attributes from Al-Maani and Bani Salameh (2017) is not explicitly defined. 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. This score is taken "as is" to determine if AI systems can functionally emulate this human benchmark in the context of shortlisting top LMS contenders.

All three AI systems provided complete scores for 97 of the 100 LMSs, which served as the final dataset for this analysis. For each AI system, the usability factor score was derived by extracting the first principal component via Principal Component Analysis. In the aforementioned prior work, they are referred to as ClFactorScore, GPFactorScore, and GeFactorScore corresponding respectively to Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash. In contrast, the Capterra's "Overall Score" was used as the human benchmark.

ANALYSIS

The primary analytical tool for this study is Rank-Biased Overlap (RBO). RBO is a similarity measure for ranked lists that is designed to be top-weighted, meaning it is more sensitive to agreement at the top of the ranking, which is often the most critical part for decision-making. Its key features are:

Handles Incomplete Lists: RBO can compare lists of different lengths and handles ties.

Top-Weighted: It includes a parameter p (persistence) ranging from 0 to 1. A lower p value gives more weight to the top ranks. For instance, $p = 0.9$ means the user is expected to look at the first $1 / (1 - p)$ positions, which is 10 for $p = 0.9$. A value of $p = 0.7$ is more focused on the very top of the list (the first 3-4 items).

Interpretable: RBO ranges from 0 (no agreement) to 1 (identical rankings).

For this analysis, RBO was computed between the following pairs of evaluators:

Inter-AI: Claude-Opus-5 vs GPT-5.6-Sol, Claude-Opus-5 vs Gemini-3.5-Flash, and GPT-5.6-Sol vs Gemini-3.5-Flash.

AI-to-Human: Each AI system's usability factor score vs the Capterra's human user aggregate score.

These comparisons were performed at four distinct depths: top 10, top 20, top 50, and the full list of 97 (i.e., top 97). The depths represent the number of top-ranked LMSs considered in the comparison.

Additionally, for each comparison and depth, three persistence parameter values $p = 0.9$, 0.8, and 0.7 were used to assess how the agreement changes when focusing more heavily on the top of the rankings. Bias-corrected and accelerated (BCa) Bootstrapped (with 2000 samples) 95% confidence intervals were also calculated for the full-list RBO estimates to provide a measure of statistical uncertainty. The analysis was performed using RBO implementation in Python. The implementation's programming was assisted by the AI system DeepSeek and cross-verified by another AI system Copilot.

RESULTS

The RBO analysis yielded several key findings regarding the ranking consistency between the AI systems and the human benchmark, the key trends being summarized below.

Inter-AI Ranking Consistency

All three AI systems demonstrated substantial ranking agreement with each other, as shown in Table 1. The RBO values were consistently high, particularly when considering the full list of LMSs and using a higher persistence parameter p.

Table 1: Inter-AI RBO values ($p = 0.9$).

Comparison	Top 10	Top 20	Top 50	Top 97
Claude-Opus-5 vs. GPT-5.6-Sol	0.796	0.800	0.639	0.535
Claude-Opus-5 vs. Gemini-3.5-Flash	0.878	0.855	0.798	0.634
GPT-5.6-Sol vs. Gemini-3.5-Flash	0.844	0.839	0.688	0.440

A clear trend emerged in that as the depth of the comparison increased from the top 10 to the full list of 97, the RBO values generally decreased. This indicates that the AI systems show stronger agreement on which LMSs are at the very top of the list, but their rankings diverge more as one considers a broader set of LMSs.

AI-to-Human Ranking Consistency

The agreement between each AI system's ranking and Capterra's human benchmark was substantially lower than the inter-AI agreement, as detailed in Table 2.

Table 2: AI-to-human RBO values ($p = 0.9$)

Comparison With Human Benchmark	Top 10	Top 20	Top 50	Top 97
Claude-Opus-5	0.661	0.481	0.279	0.128
GPT-5.6-Sol	0.714	0.486	0.287	0.196
Gemini-3.5-Flash	0.706	0.483	0.305	0.119

The RBO values are moderate for the top 10 (ranging from 0.661 to 0.714) but decrease dramatically as more LMSs are included, dropping to rather low levels (as low as 0.119) for the full list of 97 LMSs. This suggests that while the AI systems can broadly agree with humans on the very top contenders, their overall rankings are not well aligned when considering a larger set of LMSs.

Effect of Persistence Parameter (p)

The persistence parameter p controls the weight given to the top of the ranking. Among the fine-grained analysis performed for $p = 0.9$, 0.8, and 0.7 in this study, key findings can be exemplified by the Claude-Opus-5 vs. GPT-5.6-Sol comparison across the top 10 LMSs shortlisted by them as follows:

- For $p = 0.9$, RBO = 0.796
- For $p = 0.8$: RBO = 0.639
- For $p = 0.7$: RBO = 0.512

This trend is consistent across the comparisons between most AI system pairs and between most AI systems and Capterra's human benchmark: a lower p value, which focuses more heavily on the very top of the list, results in a lower RBO. This means that the agreement between an AI system pair or between an AI system and Capterra's human benchmark is weaker at the very top than when considering the two broader lists shortlisted in the comparison (e.g., the two lists of top 10 LMSs picked by Claude-Opus-5 and GPT-5.6-Sol, respectively, in the above example). This highlights the challenge for AI systems in perfectly identifying the single best LMS. Nevertheless, there are scant counter-examples. For example, the GPT-5.6-Sol vs. Gemini-3.5-Flash comparison across the full list of 97 LMSs yielded:

- For $p = 0.9$, RBO = 0.440154
- For $p = 0.8$, RBO = 0.458959
- For $p = 0.7$, RBO = 0.535515

In summary, there is no absolutely generalized trend of RBO as p varies despite their simultaneous decrease (increase) being the mainstream.

Effect of Depth

The effect of depth on RBO is also clear but mild. As the depth goes down to include more LMSs in the comparison, the RBO value generally increases very gradually. For instance, for the GPT-5.6-Sol vs. Capterra's human benchmark comparison across the full list of 97 LMSs at $p = 0.9$:

- For the depth of top 10, RBO = 0.173
- For the depth of top 20, RBO = 0.170
- For the depth of top 30, RBO = 0.191
- For the depth of top 40, RBO = 0.195
- For the depth of top 50, RBO = 0.195
- For the depth of top 60, RBO = 0.195
- For the depth of top 70, RBO = 0.195
- For the depth of top 80, RBO = 0.195
- For the depth of top 90, RBO = 0.195
- For the full depth of 97 LMSs, RBO = 0.196

This said, there could be some very miniscule fluctuations of the RBO values when descending the depth. For example, for the Claude-Opus-5 vs. Gemini-3.5-Flash comparison across the full list of 97 LMSs at $p = 0.9$, the descend experiences very minor ups and downs of RBO though increasing in general.

This indicates that the comparisons between the AI systems or between AI systems and Capterra's human benchmark give rise to rankings that are less similar for the top-tier LMSs, but they converge slightly when evaluating the majority of the LMSs.

DISCUSSION AND CONCLUSION

This study extends the findings of a prior work by shifting the focus from score correlation to ranking consistency. The results provide a more nuanced picture of the potential for AI to conduct automated usability evaluations and shortlist top-performing LMSs.

The high inter-AI ranking consistency (e.g., in the Claude-Opus-5 vs. Gemini-3.5-Flash comparison across the top 20 LMSs shortlisted by them, RBO = 0.878 at $p = 0.9$) aligns with and strengthens the prior findings. It suggests that different advanced AI systems process user feedback in a similar manner, leading to highly convergent rankings of LMSs. This reinforces the idea that there is a stable, underlying signal related to usability that these systems are extracting from textual data (Singh Chauhan, 2022). From a

systems engineering perspective, this is promising, as it suggests that employing one well-established AI system for such a task could yield reasonably reliable and replicable results.

However, the more critical finding concerns the AI-to-human ranking consistency. While the corresponding RBO values for comparison across the top 10 LMSs (up to ~0.7) are comparable to some inter-AI ranking consistency, suggesting reasonable AI-to-human agreement on the very best LMS candidates, the dramatic drop in RBO as the depth increases is a crucial limitation. For instance, the RBO values for AI-to-human comparisons across the full list of 97 LMSs (e.g., for Gemini-3.5-Flash vs. human benchmark at $p = 0.9$, RBO = 0.119) are rather low, indicating that the AI system's ranking of the majority of LMSs is not well-aligned with the human benchmark. This does not imply that AI is "wrong" in an absolute sense, but it does suggest that its evaluative criteria or the way it weighs different attributes may differ from the human users' composite judgment captured by Capterra's "Overall Score." This could be due to several factors:

1. **Noise in the Input:** The AI systems synthesize information from a vast, potentially noisy corpus of text, which may contain biases not present in Capterra's score (Liu, 2025).
2. **Differing Constructs:** Capterra's score is a measure of overall satisfaction, which may include factors beyond the seven usability attributes that the AI systems were instructed to focus on, such as price, customer support, or integrations. Conversely, Capterra's score may not fully represent all the seven attributes equally. Capterra's score was treated "as is," and the results show that AI's functional ability to replace this specific human benchmark is currently limited.
3. **Calibration:** AI's scoring and ranking methodology may be systematically calibrated differently than Capterra's score.

The observed effects of the persistence parameter p are also informative. The fact that RBO values decrease as p decreases (i.e., as one focuses more narrowly on the very top of the list) suggests that the AI systems struggle most with identifying the single "best" LMS. The agreement is stronger when considering a broader set of candidates. This could indicate that AI is better at making general qualitative distinctions than fine-grained quantitative ones.

The implications for HSI and systems engineering are significant. The results suggest that AI can be a valuable tool for rapid screening and shortlisting by usability. An HSI practitioner could, for instance, use an LLM to quickly analyze user feedback for a large number of LMSs and generate a shortlist of 10-20 viable candidates with a reasonable degree of confidence. This shortlist could then be subjected to more rigorous, traditional human evaluation. This would save significant time and resources while ensuring that no obviously poor-performing LMSs are overlooked. However, the low agreement on the full-list ranking implies that AI is not yet ready to replace human judgment for comprehensive evaluation or fine-grained ranking of a large set of LMSs (Folds et al., 2008).

Limitations and Future Work: This study has several limitations. First, the sample of LMSs, while large, may not be fully representative of all available LMSs. Second, the exact nature of Capterra's score is unknown, making direct comparisons challenging. Third, only three AI systems were examined; future work should expand to include more systems and perhaps multiple runs to assess stability. Future research should also explore the potential for fine-tuning LLMs on usability-specific datasets to improve their ranking accuracy and reduce the gap with human benchmarks (Ravikumar and Aarthi, 2026). Another promising direction is to analyze the specific attributes where AI and human rankings diverge most, which could provide valuable insights into the strengths and weaknesses of the AI systems.

In conclusion, while AI systems show promise for the functional task of shortlisting top LMS contenders by usability, their current capabilities are a supplement to, rather than a substitute for, comprehensive human evaluation. Their strength lies in rapid, scalable initial screening, which can be a powerful augmentation to the HSI toolkit. Moreover, this study's findings on rankings provide more practical utility than those on correlations in the prior work (Chan, 2026) in that in many real-world applications, such as software system selection or market analysis, rankings are a key determinant for shortlisting or even the ultimate selection decision.

ACKNOWLEDGMENT

The author would like to acknowledge the use of the AI platforms Claude-Opus-5, GPT-5.6-Sol, and Gemini-3.5-Flash for scoring LMSs' usability. 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). Consistency of AI-driven usability evaluations for software: Learning management systems as an example, Applied human factors and ergonomics (AHFE) international Volume 33 (in press).

Chapelle, O. and Zhang, Y. (2009). "A dynamic Bayesian network click model for web search ranking", *Proceedings of the 18th international conference on World wide web (WWW '09),* Madrid, Spain. https://doi.org/10.1145/1526709.1526711

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

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

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings, *ACM Transactions on Information Systems (TOIS)* Volume 28 No. 4. https://doi.org/10.1145/1852102.1852106

Yilmaz, E., Aslam, J. A. and Robertson, S. (2008). "A new rank correlation coefficient for information retrieval", *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08),* Singapore, Singapore. https://doi.org/10.1145/1390334.1390435