A Comparative Study of AI and Human Evaluation for Student Website Projects

Lidia Feklistova and Artur Kašnikov
Institute of Computer Science, University of Tartu, Estonia

lidia.feklistova@ut.ee
artur.kasnikov@ut.ee

Abstract: Artificial intelligence (AI) tools based on large language models (LLMs) are increasingly being adopted across a wide range of fields, including higher education. Given the substantial workload often faced by educators, these tools offer promising potential to assist in the evaluation of student work. However, empirical research on their reliability—particularly in assessing practical, design-oriented assignments such as student-developed websites—remains limited. This study aimed to investigate the ability of various AI tools to evaluate student website projects and the consistency between the evaluations given by AI tools and human instructors (HIs) using the same criteria. Based on a literature review, a set of evaluation criteria was developed across three categories: user interface (UI), user experience (UX), and code quality. Each student project included a website prototype and the corresponding implementation code. Nine student projects were evaluated independently by seven AI tools and HI, using a Likert scale. To reduce variability, all AI tools were provided with the same evaluation prompt. The Wilcoxon signed-rank test revealed no statistically significant differences in many evaluation criteria between AI tools and HIs, suggesting general similarity in overall scoring. On the other hand, the Spearman correlation analysis revealed low consistency in how AI tools and HI evaluated specific aspects of the projects. This indicates that while the evaluation provided by AI tools and HIs may appear similar at a surface level, their underlying judgment patterns—particularly regarding certain criteria of UI/UX design and code quality—can diverge. However, ChatGPT-4.5 and ChatGPT-4o delivered particularly promising outcomes. From an educational perspective, the study results highlight the importance of treating AI tools as supportive assistants rather than autonomous evaluators—at least for now—especially in domains involving subjective or context-sensitive judgment. Identifying where AI tools' evaluations align or conflict with human judgment provides valuable insight into the appropriate use, potential, and limitations of such tools in academic evaluation.

Keywords: Website development, UI/UX design, Code quality, Artificial intelligence tools, Large language model, Automatic vs human evaluation

1. Introduction

Artificial intelligence (AI) tools powered by large language models (LLMs) are increasingly applied across various domains. These tools enhance the accuracy of healthcare diagnostics (McKinney et al., 2020), enable the generation and intelligent editing of multimedia content (Shao et al., 2024), and support the development of reliable web applications (Ghai et al., 2024).

AI tools have also shown promise in evaluating student work. Despite some limitations in the evaluation of calculus problem solutions (Gandolfi, 2025), studies have demonstrated that AI-generated evaluations closely correspond with those of human instructors (HIs) when evaluating essays (Impey et al., 2025; Liew and Tan, 2024) and introductory programming tasks (Cisneros-González et al., 2025). Although a few studies have explored the application of LLMs in evaluating user experience (Hsueh et al., 2024), the potential of these tools for the holistic evaluation of websites remains underexplored.

This study examines the evaluation of website projects from a development perspective, encompassing both website design and coding. Website design includes user interface (UI) and user experience (UX), which are guided by well-established design principles. Coding involves the application of best practices to ensure code quality.

The study aims to investigate the ability of AI tools to evaluate website projects and the consistency between the evaluations of AI tools and HIs across different stages of website development. The following research questions are posed:

RQ1: What AI tools are able to evaluate students' website projects?

RQ2: Are there significant differences in website project evaluations provided by AI tools and HIs using the same evaluation criteria?

RQ3: Is there a correlation between website project evaluations provided by AI tools and HIs using the same evaluation criteria?

Understanding where AI tools' evaluations align or diverge from human judgment helps clarify their appropriate role and limitations in academic settings.

1.1 Web Design Principles and Code Quality

In today's digital environment, most companies maintain a website to ensure online visibility and accessibility. Visually appealing websites play a crucial role in shaping users' perceptions of trust and credibility (Arizal et al., 2024). This requires thoughtful design of both the UI and UX. UI design focuses on the aesthetics of the visual elements users interact with, including website visual style and layout. A well-considered UI design enables users to navigate through the website easily, find desired items, and complete transactions with minimal effort (Kurniawan, 2025). Such an approach contributes to a positive UX (Jansson et al., 2022), which is focused on ensuring the website performs effectively from the user's perspective (Kumar et al., 2023; Novák et al., 2024).

While numerous factors can be considered in the website UI design process, the authors of the current study focused on those they identified as most essential. In UI design, *colours* help evoke emotions, direct users' attention, establish visual hierarchy within the website (Kurniawan, 2025) and enhance users' sense of trust and satisfaction (Cyr et al., 2010). Proper *contrast* guides attention toward key elements of a website (Olejnik-Krugły et al., 2021) and makes a website more attractive to users (Jongmans et al., 2022). Carefully selected *typography* ensures readability, clarifies messaging and fosters deeper user interaction (Kurniawan, 2025). G*rouping* visual elements effectively helps users process website content more efficiently (Xiao et al., 2024).

Although UX is a multidimensional construct (Jongmans et al., 2022), this study focuses specifically on *usability*. As a key component of UX design, usability refers to how effectively, efficiently, and satisfactorily users can interact with a website (Kumar et al., 2023; Novák et al., 2024). High usability ensures that users can complete their tasks with minimal effort and confusion. Usability is directly enhanced by visual design, which supports better orientation, task flow, and overall interaction quality (Jongmans et al., 2022).

Websites' design and functionality are implemented through code, using various web development technologies. Keeping a website up and running requires continuous work on the code. This highlights the importance of code quality, which in this study is explored through two essential aspects. Code *readability* refers to how clear and accessible the source code is for reading and understanding (Martinović and Rozić, 2024; Tashtoush et al., 2023). Code readability is closely related to code *maintainability* (Martinović and Rozić, 2024; Tashtoush et al., 2023). The latter ensures that further modifications to the code can be done efficiently and without affecting the overall functionality of the website (Martinović and Rozić, 2024).

1.2 AI Tools in the Evaluation of a Website

Shao et al. (2024) argue that AI tools are able to analyse pictures. While AI can help detect layout structure and object relationships in visual media (Mishra, 2023), the specific application of AI tools to evaluate compliance with established web design principles, such as colour, contrast or usability, remains largely unexplored.

AI tools are widely used in assessing code quality (Almeida et al., 2024; Shao et al., 2024; Yi et al., 2024). These studies often treated code quality as a broad construct, typically analysing it based on overall functionality or security. However, they have not explored code readability or maintainability, which would provide a deeper understanding of how AI tools' evaluations align with those of HI.

2. Methodology

2.1 Context of Study

In this study, the analysed website projects were final assignments for the university course *Web Page Creation for Advanced Users*—an 8-week, 3 ECTS Estonian-language course. Aimed at students interested in web development, the course required no prior experience and was delivered fully online, with no in-person sessions. Learning materials, including text-based tutorials with code snippets, short videos, and supplementary links, were accessible via the university's Moodle platform from the start. The course covered six core topics, including UI/UX design principles and the use of AI in web development. Students were introduced to key technologies such as HTML, CSS, JavaScript, and Vue.js. Throughout the course, students completed practical tasks graded on a pass/fail basis by the course instructor.

For the final assignment, students applied their skills to create a comprehensive website project. Teams of two or three students self-organised, selected their own topics, and first created a multi-page *website prototype* using a design platform like Figma. Another team used this prototype to develop the *actual website* using the technologies taught.

2.2 Participants and Instruments

Although 21 students participated in the course during the autumn 2024/2025 academic year, only 18 provided consent for their projects to be analysed using AI tools. The sample consisted of an equal number (9) of female and male students. Less than one-third (28%) were master's students, while the remainder were bachelor's students. The vast majority (83%) were enrolled in the Department of Computer Science. In terms of prior web development experience, 39% reported being "somewhat satisfied" and another 39% were "neither satisfied nor dissatisfied". The remainder were "somewhat dissatisfied". As the project was completed in teams, this study analyses nine website projects. The topics of the websites varied, ranging from service platforms to online stores.

To answer the first research question, the authors selected nine different AI tools based on their personal interests. Each tool was asked to self-describe its purpose of use and its ability to work with website projects. To eliminate potential biases in the responses, this part of the research was done on the same day (April 7, 2025) using the same prompt in Estonian. This language was chosen because all evaluated website projects were in Estonian. For the purposes of this article, the original prompt has been translated into English to ensure clarity and accessibility for an international audience:

> *"I would like to learn more about the [AI tool]. Please write answers to the following questions:*
>
> *What is [AI tool] primarily intended for?*
>
> *What are the general strengths and weaknesses of [AI tool]?*
>
> *Can [AI tool] be used to evaluate websites where adherence to user interface and user experience design principles (such as typography, grouping, usability) and code quality play an important role?*
>
> *Justify your answer. Limit your answer to 5–6 sentences. Write the answer in academic Estonian."*

To ensure a theoretically grounded and pedagogically meaningful evaluation, a rubric was created based on criteria identified in the literature. These covered key principles of UI design (colour, contrast, typography, and grouping), UX design (usability), and code quality (readability and maintainability). The fulfillment of each criterion was evaluated on a Likert scale from 5 ("strongly agree") to 0 ("strongly disagree"). Two HIs and various AI tools independently evaluated all nine website projects using the created rubric. Human evaluation was based on a consensus between the authors of the current study. To evaluate UI/UX design, each AI tool received webpage screenshots; to evaluate code quality, the full source code was submitted to the AI tools. A single prompt, translated into English for this article, was used consistently across all tools to ensure comparability and reduce variability. The prompt to evaluate the website prototype was as follows:

> *"I will be uploading various screenshots of the same website project (not full-stack, but only front-end). You have to evaluate the whole project based on two design approaches: user interface (UI) and user experience (UX). Each design has its own criteria:*
>
> *User Interface (UI)*
>
> *Colour*
>
> *Contrast*
>
> *Typography*
>
> *Grouping*
>
> *User Experience (UX)*
>
> *Usability*
>
> *For each project as a whole, please evaluate the fulfillment of each criterion on a Likert scale, where 5 refers to "strongly agree" and 0 refers to "strongly disagree"".*

The prompt to evaluate the code of the developed website was as follows:

> *"I will be uploading the website project's full source code (not full-stack, but only front-end). You have to evaluate how well the code of the project is written. Use the following evaluation criteria:*
>
> *Readability*

Maintainability

For each project as a whole, please evaluate the fulfillment of each criterion on a Likert scale, where 5 refers to "strongly agree" and 0 refers to "strongly disagree"".

2.3 Data Analysis

The IBM SPSS Statistics 30 software package was used for the statistical analysis. Descriptive statistics, such as the minimum, maximum, and median, were calculated. The Shapiro-Wilk test revealed that, in many cases, the data (evaluations provided by HIs and AI tools) were not normally distributed. In addition, due to the ordinal nature of the data and the small sample size, the use of non-parametric tests was justified (Gust and D'journo, 2015; Harris et al., 2008; Schober et al. 2018). Since the two datasets included the evaluations of the same website projects, the Wilcoxon signed-rank test was employed to determine whether there were significant differences between the evaluations of AI tools and HIs. The Spearman rank correlation was employed to assess the consistency between evaluations by AI tools and HIs. According to Schober et al. (2018), correlations below 0.4 should be considered "weak" and those above 0.7 should be considered "strong".

3. Results

3.1 AI Tools' Ability to Evaluate Website Projects

In this study, nine AI tools based on various LLMs were selected to investigate their ability to evaluate students' website projects. Due to space limitations, the full-text responses of the AI tools to the prompt described in section 2.2 are not included here. However, Table 1 provides a summary of the main findings.

Table 1: Summary of AI tools' self-description (as of April 7, 2025)

AI tool	Intended use	General strengths	General weaknesses	Ability to evaluate UI/UX design principles and code quality
ChatGPT -4o	Natural language processing	Context awareness, fluency, variety of knowledge	Factual accuracy, data freshness	Can be applied to website project evaluation, especially when assessing UI design compliance with general principles; can analyse code quality in general. *
ChatGPT -o1	Text-based tasks solving	Context awareness, fluency, variety of knowledge	Factual inaccuracy, nuance handling	Can provide general recommendations on UI design, but critical analysis of finer design details may be limited; to some extent, may assess code quality and semantic inconsistencies. *
ChatGPT -4.5	Intelligent text processing	Knowledge breadth, coherence, comprehension	Factual errors, nuance limits, data dependency	Lacks deep visual and interactive understanding of user interfaces, assessing design and usability indirectly through text-based recommendations; in code, can identify standard and clearly structured solutions but may lack an understanding of complex code nuances. *
Gemini 2.0 Flash	Low-latency response generation	Speed, processing power	Size limit, analytical weakness	Lacks contextual understanding and aesthetic analysis to assess UI design principles carefully; can somewhat analyse code structure and detect patterns. *
Gemini 2.0 Pro	Complex tasks with multimodal processing	Big data processing, context handling	Hallucination, bias caused by training data	Lacks subjective experience and the ability to evaluate actual user experience and visual appeal; can be used to analyse code quality in website projects by detecting errors, assessing style guideline adherence, and evaluating code complexity. *
Claude 3.7 Sonnet	Natural language understanding and generation	Knowledge breadth, context understanding, detailed answers	Outdated knowledge, novelty barrier	Can analyse both UI/UX design principles and code quality because it is trained to understand design concepts and code structures; the evaluation of code relies on the input provided, but cannot independently visit websites.
Mistral	Open-source LLM creation, customization	Flexibility, performance, low latency	Text-oriented, visual incapacity	Can offer usability improvement suggestions, but visual design principles are not its main strengths; can assist in analysing code quality. *

AI tool	Intended use	General strengths	General weaknesses	Ability to evaluate UI/UX design principles and code quality
Deep-Seek	Complex text-based tasks solving	Context understanding, problem-solving, ability of analysis	Visual assessment limitations	Can analyse UI design principles based on descriptions; can assess code quality for readability, structure, and standards compliance. *
Copilot	Natural language process assistance	Natural language generation, context awareness	Cultural insensitivity, bias caused by training data	Can evaluate website projects by analysing and providing suggestions on UI design principles; however, its ability is based on statistical analysis of texts; can evaluate code quality.

Many AI tools (marked with an asterisk in Table 1) emphasised that they are best suited for providing initial analysis and recommendations, but not for final quality assessment in technically complex projects. Based on their responses, these tools suggest being used as assistants, with the final evaluation performed by a human. DeepSeek and Copilot, which only provide feedback on text descriptions, were excluded from further analysis.

3.2 Comparison of Website Project Evaluations Provided by HI and Various AI Tools Based on Same Criteria

The authors of the current study analysed nine website projects that were evaluated by two HIs and seven AI tools using the same rubric described in section 2.2. The results for each criterion are provided in Table 2, where z- and p-values correspond to the Wilcoxon signed-rank test results. Statistically significant differences between HI and AI evaluations are highlighted in bold.

Table 2: Comparison of website project evaluations

	HIs	ChatGPT-4o	ChatGPT-o1	ChatGPT-4.5	Gemini 2.0 Flash	Gemini 2.0 Pro	Claude 3.7 Sonnet	Mistral
Colour								
Mean	4.22	3.44	3.33	4.11	4.11	3.44	3.22	3.44
SD	.972	.882	.500	1.054	.601	.726	.667	.527
Min-Max	3-5	2-4	3-4	2-5	3-5	3-5	2-4	3-4
Median	5	4	3	4	4	3	3	3
z-value		-2.646	-2.070	-.447	-.302	-1.838	-2.081	-2.070
p-value		**.008**	**.038**	.655	.763	.066	**.037**	**.038**
Contrast								
Mean	3.67	3.89	3.44	4.44	3.89	4.44	3.00	3.11
SD	.866	.928	.726	.882	.333	.527	.866	.333
Min-Max	3-5	3-5	2-4	3-5	3-4	4-5	2-4	3-4
Median	3	4	4	5	4	4	3	3
z-value		-.707	-.649	-1.890	-.707	-1.732	-1.730	-1.518
p-value		.480	.516	.059	.480	.083	.084	.129
Typography								
Mean	4.00	3.67	3.22	4.56	4.00	4.56	3.00	3.44
SD	1.118	.707	.441	.882	.500	.527	.500	.527
Min-Max	2-5	3-5	3-4	3-5	3-5	4-5	2-4	3-4
Median	4	4	3	5	4	5	3	3
z-value		-.750	-1.734	-1.890	-.175	-1.406	-2.121	-1.406
p-value		.453	.083	.059	.861	.160	**.034**	.160

	HIs	ChatGPT-4o	ChatGPT-o1	ChatGPT-4.5	Gemini 2.0 Flash	Gemini 2.0 Pro	Claude 3.7 Sonnet	Mistral
Grouping								
Mean	4.11	4.11	2.89	4.56	4.00	4.78	4.11	4.00
SD	1.054	.928	.601	1.014	.500	.441	.333	.000
Min-Max	2-5	2-5	2-4	2-5	3-5	4-5	4-5	4-4
Median	4	4	3	5	4	5	4	4
z-value		.000	-2.309	-1.633	-.302	-1.857	.000	-.333
p-value		1.000	**.021**	.102	.763	.063	1.000	.739
Usability								
Mean	4.22	3.89	3.56	4.67	4.11	4.44	3.44	4.00
SD	1.093	.782	.527	.707	.333	.527	.527	.000
Min-Max	2-5	3-5	3-4	3-5	4-5	4-5	3-4	4-4
Median	5	4	4	5	4	4	3	4
z-value		-1.342	-1.387	-2.000	-.176	-.378	-1.588	-.632
p-value		.180	.165	**.046**	.860	.705	.112	.527
Readability								
Mean	3.89	4.22	4.00	5.00	4.00	3.78	3.44	4.00
SD	.601	.441	.000	.000	.000	.972	.527	.000
Min-Max	3-5	4-5	4-4	5-5	4-4	2-5	3-4	4-4
Median	4	4	4	5	4	4	3	4
z-value		-1.342	-.577	-2.640	-.577	-.378	-1.414	-.577
p-value		.180	.564	**.008**	.564	.705	.157	.564s
Maintainability								
Mean	3.22	3.11	4.00	5.00	4.11	2.78	2.78	4.11
SD	.833	.601	.000	.000	.333	1.093	.833	1.054
Min-Max	2-4	2-4	4-4	5-5	4-5	1-4	2-4	3-5
Median	3	3	4	5	4	3	3	5
z-value		-.378	-2.070	-2.701	-2.060	-1.054	-.964	-1.725
p-value		.705	**.038**	**.007**	**.039**	.292	.340	.084

SD – standard deviation

Although the median values varied, the study found that overall evaluations by AI tools and HIs were not significantly different. However, significant differences appeared in some specific criteria. For example, ChatGPT-4o, ChatGPT-o1, Claude 3.7 Sonnet, and Mistral were more critical when evaluating the use of colour, awarding lower scores. In contrast, ChatGPT-o1, ChatGPT-4.5, and Gemini 2.0 Flash evaluated code maintainability higher than HIs did. Overall, evaluations by ChatGPT-4.5 tended to show greater differences from HI evaluations compared to the other studied AI tools.

3.3 Correlation Between Evaluations of Website Projects Provided by HI and Various AI Tools

The Spearman correlation was used to compare the consistency between evaluations provided by HIs and AI tools that applied the same criteria. Table 3 provides a summary of the findings.

Table 3: Correlation between HI and AI tools evaluations

Criterion		ChatGPT-4o	ChatGPT-o1	ChatGPT-4.5	Gemini 2.0 Flash	Gemini 2.0 Pro	Claude 3.7 Sonnet	Mistral
Colour	HIs	**.902*** **	.306	.644	-.122	-.111	.021	.484
Contrast	HIs	.414	-.144	.325	-.227	-.431	.321	-.303
Typography	HIs	-.005	-.327	**.709***	-.625	.137	.337	.183
Grouping	HIs	**.833****	.355	.583	.000	.551	-.146	N/A
Usability	HIs	**.819****	.048	**.804****	-.452	.191	.191	N/A
Readability	HIs	.062	N/A	N/A	N/A	.331	-.207	N/A
Maintainability	HIs	.176	N/A	N/A	-.512	-.357	-.248	-.046

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; N/A correlation could not be calculated because one of the variables had constant values; grey cell – weak correlation

There were some statistically significant strong positive correlations between project evaluations provided by HIs and studied AI tools. However, in many cases, the correlation was weak and not statistically significant. ChatGPT-4o tended to show greater consistency with HIs evaluations than other AI tools studied.

4. Discussion

The study explored the ability of AI tools to evaluate website projects. Firstly, the authors selected nine AI tools and explored their self-descriptions. The results demonstrated that ChatGPT-4o, ChatGPT-o1, ChatGPT-4.5, Gemini 2.0 Flash, Gemini 2.0 Pro, Claude 3.7 Sonnet and Mistral stated their ability to evaluate website projects. DeepSeek and Copilot pointed out the difficulties in assessing visual content, as their analysis is based on text input.

Next, nine student projects, focusing on key web design and coding principles, were evaluated using the same rubric criteria. The evaluations were provided by seven AI tools and two HIs. Results showed that ChatGPT-4o, ChatGPT-o1, Claude 3.7 Sonnet, and Mistral evaluated colour use more strictly than HI. ChatGPT-o1 gave a significantly lower evaluation for grouping, which is part of the visual hierarchy. Since aesthetics is subjective, these differences may be due to AI training data where less attention may be put on aesthetic aspects. Helliwell (2024) also noted that aesthetic judgments often differ between AI and humans.

The results showed that AI tools' evaluations of technical UI aspects such as typography and contrast were not significantly different from those of HIs. In their evaluations, AI tools can rely on the Web Content Accessibility Guidelines, which provide suggestions for typography and contrast to ensure content readability. The only significantly lower typography evaluation in comparison to HI evaluations was by Claude 3.7 Sonnet. This discrepancy is particularly interesting given that, in its self-description, Claude 3.7 Sonnet stated that it was trained to understand design concepts.

Hsueh et al. (2024) found that ChatGPT-4 generally performs better than human evaluators across various UX design dimensions. However, the current study mainly showed no significant difference between AI tools and HIs in evaluating usability. Only ChatGPT-4.5 evaluated usability significantly higher than HI. ChatGPT-4.5 might have missed some usability nuances in the prototype screenshots because, according to its self-description, it evaluates usability indirectly using text-based recommendations.

There was no significant difference between AI tools and HIs in evaluating code readability (except ChatGPT-4.5). Meanwhile, ChatGPT-o1, ChatGPT-4.5, and Gemini 2.0 Flash evaluated code maintainability higher than HI. Although Yi et al. (2024) found that AI-based systems can accurately recognise problematic patterns in code, the results of the current study suggest some AI tools may overlook nuances such as code consistency and confusing naming.

Finally, the consistency between AI tools' and HIs' evaluations was examined. The correlation test revealed a strong statistically significant positive relationship between ChatGPT-4o and HIs for website prototype colour, grouping, and usability, and between ChatGPT-4.5 and HIs for typography and usability. These results indicate a consistency in agreement on evaluating certain criteria. The findings suggest that ChatGPT-4o and ChatGPT-

4.5 can closely reproduce human evaluations for the aforementioned criteria. Cisneros-González et al. (2025) also found a strong positive correlation between grades provided by ChatCPT-4o and HIs in evaluating programming tasks. However, many correlations in the current study were weak, indicating possible inconsistencies between AI tools and HIs evaluations. This may be a consequence of the probabilistic nature of LLMs, and similar inconsistencies were reported in previous studies (Emirtekin, 2025).

5. Conclusions

This study had two aims: to assess AI tools' ability to evaluate website projects and to examine how consistently their evaluations align with those of HIs. While DeepSeek and Copilot as of April 7th, 2025 could not analyse visual input, other tools were tested on UI/UX design and code quality criteria. The Wilcoxon signed-rank tests showed few significant differences, but the Spearman correlations revealed inconsistent agreement in many cases. This suggests that, despite surface similarities, AI tools and human evaluations may differ on specific design and code elements. Therefore, AI tools should be seen as supportive assistants, not standalone evaluators, especially in tasks requiring subjective or design-sensitive judgment.

The current study is limited by a small sample size and the subjectivity of HIs' evaluations. At the time of writing this paper, the studied AI tools could not access live website prototypes or fully assess complex visuals. As AI tools continue to evolve, future research would be useful to explore their improved abilities, including providing written feedback in evaluating various types of creative works. Additionally, there is a growing case for investigating bespoke LLM development tailored to educational institutions. Future research could examine how fine-tuned or custom-trained models might better align with specific assessment goals, or institutional values.

Ethics statement: Prior to the beginning of the study, all participants received comprehensive information about it. Informed consent was obtained from all students, and their privacy was ensured throughout the study. Personal data was anonymised, and any information that could lead to student identification was carefully withheld throughout the research.

AI declaration: Studied AI tools were used to evaluate students' website projects. To enhance the readability of this paper, the authors utilised Grammarly and ChatGPT-4.5. All content was subsequently reviewed and edited by the authors to ensure accuracy and appropriateness. The authors bear full responsibility for the content presented in the final version of this paper.

References

Almeida, Y., Albuquerque, D., Filho, E.D. et al. (2024) "AICodeReview: Advancing code quality with AI-enhanced reviews", *SoftwareX*, Vol 26. https://doi.org/10.1016/j.softx.2024.101677

Arizal, N., Nofrizal, Listihana W.D. and Hadiyati (2024) "Gen Z Customer Loyalty in Online Shopping: An Integrated Model of Trust, Website Design, and Security", *Journal of Internet Commerce*, Vol 23, Issue 2, pp 121-143. https://doi.org/10.1080/15332861.2024.2330812

Cisneros-González, J., Gordo-Herrera, N., Barcia-Santos, I., and Sánchez-Soriano, J. (2025) "JorGPT: Instructor-Aided Grading of Programming Assignments with Large Language Models (LLMs)", *Future Internet*, Vol 17, Issue 6. https://doi.org/10.3390/fi17060265

Cyr, D., Head, M. and Larios, H. (2010) "Colour appeal in website design within and across cultures: A multi-method evaluation", *International Journal of Human-Computer Studies*, Vol 68, Issues 1–2, pp 1-21. https://doi.org/10.1016/j.ijhcs.2009.08.005

Emirtekin, E. (2025) "Large Language Model-Powered Automated Assessment: A Systematic Review", *Applied Sciences*, Vol 15, Issue 10. https://doi.org/10.3390/app15105683

Gandolfi, A. (2025) "GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions", *International Journal of Artificial Intelligence in Education*, Vol 35, pp 367–397. https://doi.org/10.1007/s40593-024-00403-3

Ghai, A.S., Rawat, V., Gupta, V.K. and Ghai, K.P. (2024) "Artificial Intelligence in System and Software Engineering for Auto Code Generation", *International Conference on Electrical Electronics and Computing Technologies*. Greater Noida, India, pp 1-5. https://doi.org/10.1109/ICEECT61758.2024.10738945

Gust, L. and D'Journo, X.B. (2015) "The use of correlation functions in thoracic surgery research", *Journal of thoracic disease*, Vol 7, No 3. http://doi.org/10.3978/j.issn.2072-1439.2015.01.54

Harris, J.E., Boushey, C., Bruemmer, B. and Archer, S.L. (2008) "Publishing Nutrition Research: A Review of Nonparametric Methods, Part 3", *Journal of the American Dietetic Association*, Vol 108, Issue 9, pp 1488-1496. https://doi.org/10.1016/j.jada.2008.06.426

Helliwell, A.C. (2024) "Aesthetic Value and the AI Alignment Problem", *Philosophy & Technology*, Vol 37. https://doi.org/10.1007/s13347-024-00816-x

Hsueh, N-L., Lin, H-J., Lai, L-C. (2024) "Applying Large Language Model to User Experience Testing", *Electronics*, Vol 13, Issue 23. https://doi.org/10.3390/electronics13234633

Impey, C., Wenger, M., Garuda, N., Golchin, S. and Stamer, S. (2025) "Using Large Language Models for Automated Grading of Student Writing about Science", *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00453-7

Jansson, M., Liisanantti, J., Ala-Kokko, T. and Reponen, J. (2022) "The negative impact of interface design, customizability, inefficiency, malfunctions, and information retrieval on user experience: A national usability survey of ICU clinical information systems in Finland", *International Journal of Medical Informatics*, Vol 159. https://doi.org/10.1016/j.ijmedinf.2021.104680

Jongmans, E., Jeannot, F., Liang, L., and Dampérat, M. (2022) "Impact of website visual design on user experience and website evaluation: The sequential mediating roles of usability and pleasure", *Journal of Marketing Management*, Vol 38, Issues 17-18, pp 2078-2113. https://doi.org/10.1080/0267257X.2022.2085315

Kumar, V., Kumar, V., Singh, S., Singh, N. and Banoth, S. (2023) "The impact of user experience design on customer satisfaction in e-commerce websites", *International Journal for Research in Applied Science and Engineering Technology*, Vol 11, Issue 5, pp 4571–4575. https://doi.org/10.22214/ijraset.2023.52580

Kurniawan, A. (2025) "The impact of golden ratio application on user satisfaction: A study on horizontal scrolling in user interface (UI) design". *International Journal of Human–Computer Interaction*, Vol 41, Issue 1, pp 445–451. https://doi.org/10.1080/10447318.2023.2301254

Liew, P.Y. and Tan., I. K. (2024) "On automated essay grading using large language models", *8th international conference on computer science and artificial intelligence*. Beijing, China, pp 204-211. https://doi.org/10.1145/3709026.3709030

Martinović, B., and Rozić, R. (2024) "Impact of AI tools on software development code quality", *Digital transformation in education and artificial intelligence application*. Mostar, Bosnia and Herzegovina, pp 241-256. https://doi.org/10.1007/978-3-031-62058-4_15

McKinney, S.M., Sieniek, M., Godbole, V. et al. (2020) "International evaluation of an AI system for breast cancer screening", *Nature*, Vol 577, pp 89-94. https://doi.org/10.1038/s41586-019-1799-6

Mishra, S. (2023) "An Analysis of How Artificial Intelligence is Used in the Field of Image Identification", *Journal for Research in Applied Sciences and Biotechnology*, Vol 2, Issue 3, pp 106-113. https://doi.org/10.55544/jrasb.2.3.14

Novák, J.Š., Masner, J., Benda, P., Šimek, P., and Merunka, V. (2024) "Eye Tracking, Usability, and User Experience: A Systematic Review", *International Journal of Human–Computer Interaction*, Vol 40, Issue 17, pp 4484-4500. https://doi.org/10.1080/10447318.2023.2221600

Olejnik-Krugły, A., Tomaszewska, A., Dziśko, M., and Jankowski, J. (2021) "Towards effective visual communication with positive user experience: High contrast and visibility vs. userfriendliness/positive perception".

Schober, P., Boer, C. and Schwarte, L.A. (2018) "Correlation Coefficients: Appropriate Use and Interpretation", *Anesthesia & Analgesia*, Vol 126, No 5, pp 1763-1768. https://doi.org/10.1213/ane.0000000000002864

Shao, M, Basit, A., Karri R. and Shafique, M. (2024) "Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges", *IEEE Access*, Vol 12, pp 188664-188706. https://doi.org/10.1109/ACCESS.2024.3482107

Shuhong X., Chen, Y., Song, Y. et al. (2024) "UI semantic component group detection: Grouping UI elements with similar semantics in mobile graphical user interface", *Displays*, Vol 83. https://doi.org/10.1016/j.displa.2024.102679

Tashtoush, Y., Abu-El-Rub, N., Darwish, O., Al-Eidi, S., Darweesh, D., and Karajeh, O. (2023) "A Notional Understanding of the Relationship between Code Readability and Software Complexity", *Information*, Vol 14, Issue 2. https://doi.org/10.3390/info14020081

Yi, S., Yu, Y. and Wu, J. (2024) "AI-based Online Code Quality Assessment System", *3rd International Conference on Cloud Computing, Big Data Application and Software Engineering.* Hangzhou, China, pp. 659-663. https://doi.org/10.1109/CBASE64041.2024.10824380