UXMind: An Intelligent Agent for AI-Assisted Interface Evaluation in UX Education

Danielle Pompeu Noronha Pontes, State University of Amazonas, Brazil

The Paris Conference on Education 2026
Official Conference Proceedings

Abstract

Interface evaluation methods play a central role in the field of Human–Computer Interaction (HCI), enabling the systematic analysis of usability, accessibility, and the visual organisation of interactive systems. However, applying these methods in educational contexts often requires methodological guidance and analytical experience from students. This paper presents UXMind, an intelligent agent designed to support AI-assisted interface evaluation in user experience (UX) education. The agent, available at https://danipontes.github.io/uxmind/, integrates multiple theoretical frameworks into a single web-based platform, enabling structured evaluations and the automated generation of interpretive analyses. UXMind operationalises 47 evaluation criteria derived from four widely recognised frameworks in the HCI literature: Nielsen's usability heuristics, Bastien and Scapin's ergonomic criteria, the Web Content Accessibility Guidelines (WCAG 2.1), and Gestalt principles applied to visual perception. During the evaluation process, the agent perceives the evaluation environment by collecting the scores assigned to each criterion and the qualitative observations recorded by evaluators. These data are processed by a Large Language Model (LLM) that acts as the cognitive component of the agent, enabling it to reason over usability evidence, identify evaluation patterns, and produce structured interpretations. The interpretive analysis is generated through an LLM accessed via the Groq API, a high-performance inference platform designed for low-latency execution of language models. By integrating interface evaluation frameworks with generative artificial intelligence and intelligent agent capabilities, UXMind contributes a novel approach to AI-assisted UX education, enabling scalable interface inspection practices and supporting the development of critical evaluation skills in HCI training.

Keywords: heuristic evaluation, generative artificial intelligence, HCI education, large language models (LLMs), user experience (UX)

iafor
The International Academic Forum
www.iafor.org

Introduction

The growing digitalisation of public, commercial, and educational services has established interface design as a strategic competency in the training of Information Technology professionals. The quality of user experience (UX) has become, in this context, a factor with implications that extend well beyond the commercial performance of digital products: poorly designed interfaces can function as effective barriers to access for populations with varying levels of digital literacy, motor abilities, or accessibility needs (Henry et al., 2014; W3C, 2024).

The COVID-19 pandemic, which emerged in 2020, acted as a catalyst for global digital transformation. Research by McKinsey & Company (LaBerge et al., 2020) estimated that the digitalisation of products and services was accelerated by approximately seven years as a result of the disruptions imposed by social distancing measures. This shift brought into sharp relief the relevance of interface quality, making UX a competitive differentiator of the first order (Soto-Acosta, 2020). The global UX market, valued at USD 9.96 billion in 2024, is projected to reach USD 33.19 billion by 2032 (Verified Market Research, 2025). Over the same period, the U.S. Bureau of Labor Statistics projects 8% growth in employment in web development and digital interface design between 2023 and 2033—a rate above the general occupational average. This picture is further reinforced by the finding that 71% of UX professionals believe Artificial Intelligence will transform design in the coming years, with automation and predictive design gaining increasing prominence (UXness Survey, 2024).

This context highlights the need for undergraduate Computer Science programmes to prepare professionals who are not only capable of applying interface evaluation methods, but of interpreting their findings and communicating them with appropriate technical grounding.

In the field of Human-Computer Interaction (HCI), heuristic evaluation, proposed by Nielsen and Molich (1990) and consolidated by Nielsen (1994), is widely adopted for its efficiency: it requires no direct participation of end users and allows usability problems to be identified in a systematic manner. Relying exclusively on a single framework, however, tends to limit diagnostic coverage. Cockton and Woolrych (2001) observed that different inspection methods capture distinct categories of problems and that the overlap between findings from different frameworks is only partial, suggesting the need to combine them. Added to this is the empirical evidence presented by Nielsen (1994): individual evaluators identify, on average, only 35% of an interface's problems, and the detection rate grows logarithmically with the number of evaluators, reaching approximately 75% with five specialists. These findings point to the relevance of instruments that enable the consolidation of evaluations conducted by multiple reviewers.

In parallel with this context, generative artificial intelligence (GenAI) has emerged as one of the most consequential phenomena in higher education. Systematic reviews published between 2023 and 2024 document substantial growth in the adoption of GenAI in educational settings, spanning from instructional strategies to automated assessment and feedback generation (Li et al., 2025). In the specific domain of interface evaluation, Hsueh et al. (2024) indicated that LLM-based approaches may reduce the costs of the evaluation process without compromising diagnostic quality relative to traditional methods. Zhong et al. (2025), in turn, found that synthetic LLM-based systems detected between 73% and 77% of usability problems—a rate higher than that observed among individual human evaluators, which ranged from 57% to 63%.

Despite these advances, a systematic review of GenAI integration in UI/UX workflows, covering the period from 2020 to 2025, found that the interface evaluation stage remains relatively underexplored compared to prototyping and visual generation (Kumar et al., 2025). This gap is even more pronounced in pedagogical contexts: tools that combine multidimensional heuristic evaluation, collaborative classroom dynamics, and generative AI support for HCI teaching are seldom reported in the literature.

Drawing on this diagnosis, the present work addresses the following research question: how can intelligent agents based on language models support students in learning interface evaluation in UX education, particularly with regard to the application of evaluation methods and the interpretation of their results? To answer it, this paper proposes and describes UXMind, an LLM-based intelligent agent that integrates multiple interface evaluation frameworks and generates interpretive analyses from the data collected during the evaluation process.

Literature Review

Interface Evaluation Methods

Among usability inspection methods, heuristic evaluation holds a prominent position owing to its favourable cost-to-benefit ratio (Nielsen, 1994). The ten heuristics proposed by Nielsen, Visibility of System Status, Match Between System and the Real World, User Control and Freedom, Consistency and Standards, Error Prevention, Recognition Rather Than Recall, Flexibility and Efficiency of Use, Aesthetic and Minimalist Design, Help Users Recognize, Diagnose, and Recover from Errors, and Help and Documentation, stand as the most frequently cited reference framework in usability evaluation studies published between 2010 and 2023, with applications spanning web, mobile, and virtual and augmented reality environments (Ntoa, 2025).

The ergonomic criteria of Bastien and Scapin (1993), developed at INRIA, offer a level of analytical granularity that exceeds that of Nielsen's heuristics, organising 18 criteria across eight dimensions—Guidance, Workload, Explicit Control, Adaptability, Error Management, Consistency, Significance of Codes, and Compatibility—which allows for more precise diagnosis of specific interaction problems (Nassar, 2012). The WCAG 2.1 guidelines from the W3C (2024), in turn, structure 78 success criteria around four principles—Perceivable, Operable, Understandable, and Robust (POUR)—and constitute the international technical standard for digital accessibility; their integrated evaluation alongside usability is considered indispensable in user-centred design approaches (Henry et al., 2014). The Gestalt principles (Wertheimer, 1923 as cited in Lidwell et al., 2003), finally, describe the mechanisms by which the human perceptual system groups and organises visual elements, offering evaluators conceptual tools to examine the perceptual dimension of interface design.

The combined use of these four frameworks finds empirical support in the literature: Cockton and Woolrych (2001) observed that the overlap between problems detected by distinct approaches is only partial, suggesting that evaluations conducted from multiple perspectives tend to broaden diagnostic coverage relative to the use of a single framework in isolation.

Generative Artificial Intelligence in Education

GenAI, driven by the development of Large Language Models (LLMs), has demonstrated the capacity to produce coherent and contextually situated content from natural language instructions (Brown et al., 2020). Giannakos et al. (2025) mapped four domains in which GenAI shows pedagogical potential: personalised instructional design, learning regulation, content generation, and assessment with feedback. Research in computer science education suggests that LLMs may offer feedback of comparable quality to that of human tutors in programming tasks, with the added advantage of immediate availability and scalability (Finnie-Ansley et al., 2023; Prather et al., 2024). In HCI teaching, the possibility of an LLM interpreting heuristic evaluation data and structuring technical reports may be understood as an extension of these capabilities to the interface design domain—though such a transposition is neither straightforward nor free of limitations.

The integration of GenAI in educational contexts, however, presents challenges that warrant attention. Giannakos et al. (2025) highlight concerns related to excessive student dependency, academic integrity, model bias, and data privacy. For LLM-integrated systems in HCI tools specifically, Navarro et al. (2026) recommend that studies explicitly justify the choice of model, characterise the expected behaviour of the LLM, and include technical evaluation complementary to user studies.

Digital Tools for UX/HCI Education

HCI education faces a challenge of a structural nature: it is an essentially practical discipline whose effective learning presupposes the application of theoretical concepts to real interfaces (Cockton et al., 2012). In this regard, collaborative heuristic evaluation—in which multiple evaluators examine the same interface and their findings are subsequently consolidated—tends to yield more comprehensive diagnoses than evaluations conducted individually (Hertzum & Jacobsen, 2001; Nielsen, 1994).

Tools that support this collective process, enabling the collection, consolidation, and comparison of evaluations carried out by multiple participants, present relevant potential both for professional practice and for educational contexts. Alomari et al. (2020), in proposing an evaluation framework for cyberlearning environments in computer science courses, observed that support for the practical learning of HCI methods tends to be insufficiently addressed in the design of educational tools for the field.

Methodology

This study adopts Design Science Research (DSR) as its guiding methodological paradigm. Proposed by Hevner et al. (2004, as cited in Peffers et al., 2007) in the field of Information Systems, DSR is grounded in the principle that knowledge about a problem and its possible solutions is constructed through the design and evaluation of purposefully built artefacts. Unlike the behavioural paradigm, which is oriented towards explaining existing phenomena, DSR focuses on the creation of innovations that extend human capabilities through the development of new artefacts—constructs, models, methods, or technological instantiations. The adoption of this approach is justified by the nature of the contribution proposed here: UXMind is conceived as a technological-pedagogical artefact that seeks to address the scarcity of tools capable of supporting students in learning interface evaluation in HCI courses.

The study followed the model proposed by Peffers et al. (2007), which organises the DSR process into six activities: (1) problem identification and motivation; (2) definition of solution objectives; (3) design and development of the artefact; (4) demonstration; (5) evaluation; and (6) communication.

Problem identification drew on a literature review conducted in the Scopus, ACM Digital Library, and Google Scholar databases, covering the period from 2015 to 2025, using the search terms heuristic evaluation tools, UX education, HCI teaching, and LLM usability evaluation. The analysis allowed three gaps to be identified: (a) the absence of pedagogical tools that bring together multiple heuristic evaluation frameworks within a single platform; (b) the lack of solutions that combine collaborative classroom evaluation dynamics with automatic consolidation of the data produced; and (c) the underexploration of the interface evaluation stage in studies on the integration of GenAI into UX/HCI workflows (Kumar et al., 2025).

Drawing on this diagnosis, functional and pedagogical requirements were established across three dimensions: technical (support for multiple frameworks; standardised scoring scales; data export); pedagogical (support for individual and collective evaluations; consolidation and visualisation of evaluations from multiple students; generation of interpretive analyses); and accessibility and portability (operation without installation or registration; compatibility with mobile devices; free browser-based access; multilingual support in Portuguese and English).

UXMind was developed through an iterative process comprising three refinement cycles, as recommended by DSR, incorporating feedback analysis obtained in real pedagogical contexts—namely, the User Experience (UX) Track courses of the Samsung Ocean R&D Project and the HCI course of the undergraduate Computer Science programme at the State University of Amazonas (UEA). The demonstration of the artefact took place in activities within these educational settings, in which students used UXMind to conduct heuristic evaluations of interfaces selected by the instructor. The preliminary evaluation combined static analysis of the tool's structure, descriptive evaluation based on use scenarios, and direct classroom observation. This article is therefore situated in the proposition and description phase of the DSR cycle; empirical evaluation with users and studies of pedagogical efficacy constitute the agenda for future work.

Results

Technical Architecture

UXMind's architecture is minimalist and deliberately portable. The entire application is encapsulated in a single HTML5 file, with no dependencies on external JavaScript frameworks—a decision driven by two central pedagogical requirements: reducing barriers to access and ensuring operation in contexts with limited connectivity, a recurring situation in public educational institutions in the Amazonian region. The AI analysis mechanism was implemented through integration with the Llama 3 model API via Groq, using a serverless proxy architecture hosted on the Vercel platform, which protects access credentials and enables free use in educational contexts. The generated analysis is transmitted to the user via Server-Sent Events (SSE), with a progressive typing effect. The technological components are summarised in Table 1.

Table 1
UXMind Technological Architecture

Component	Description and Function
HTML5 + CSS3 + JavaScript	Single-file application (SPA), without external dependencies. Ensures full portability and offline operation for evaluation functionalities.
GitHub Pages	Free and stable hosting with a permanent public URL, suitable for academic citation and student distribution.
Serverless Proxy (Vercel)	Intermediary between the user interface and the AI API. Protects access credentials and enables free use in the educational context.
Groq API / Llama 3 (LLM)	Language model used for the generation of interpretive analyses, with a free tier suitable for the academic context.
SSE Streaming	Server-Sent Events for real-time transmission of the analysis, with a progressive typing effect.
LocalStorage	Browser-based local storage for persistence of evaluations in Class Mode, without the need for an external database.
Internationalization (i18n)	Bilingual translation system (Portuguese/English) covering the interface, evaluation criteria, and AI prompts.

Source: Research Data.

Knowledge Base: Evaluation Criteria

UXMind's knowledge base comprises 47 evaluation criteria distributed across four scientific frameworks (Table 2). Each criterion is presented with a unique alphanumeric identifier, title, technical definition extracted from primary sources, examples of positive and negative application, and a scoring field on a scale from 0 to 4 (Poor to Excellent), with an N/A option for non-applicable criteria. The inclusion of technical definitions serves a dual pedagogical function: it guides the student throughout the inspection process and provides the LLM with the specialised vocabulary required to produce technically grounded analyses (Zhong et al., 2025).

Table 2
UXMind Evaluation Frameworks and Criteria

Framework	Criteria / Dimensions
Nielsen (1994)	10 usability heuristics: Visibility of System Status, Match with the Real World, User Control and Freedom, Consistency, Error Prevention, Recognition, Flexibility, Minimalist Design, Error Recovery, and Help.
Bastien and Scapin (1993)	18 ergonomic criteria across 8 dimensions: Guidance, Workload, Explicit Control, Adaptability, Error Management, Consistency, Significance of Codes, and Compatibility.
WCAG 2.1 — W3C (2024)	11 accessibility criteria under the POUR principles: Perceivable, Operable, Understandable, and Robust.
Gestalt (Wertheimer, 1923 as cited in Lidwell et al., 2003)	8 principles of visual perception: Proximity, Similarity, Continuity, Closure, Figure and Ground, Visual Hierarchy, Common Region, and Symmetry.

Source: Research Data.

Main Features

In the individual evaluation mode, a single evaluator navigates through the 47 criteria in an interface organised into thematic tabs, assigning scores and recording qualitative observations for each item. A global progress bar displays the percentage of criteria evaluated per framework—a feature directly related to Nielsen's Heuristic H01 (Visibility of System

Status). Upon completion, the student accesses a dashboard showing an overall weighted score, averages per framework, a map of critical criteria and strengths, and export options in CSV and TXT formats.

The Class Mode (Modo Turma) constitutes the tool's most pedagogically significant feature. Grounded in the literature on collaborative heuristic evaluation (Hertzum & Jacobsen, 2001; Nielsen, 1994) and technology-mediated collaborative learning (Dillenbourg, 1999), the mode operates as follows: (1) the instructor configures a session, defines the interface to be evaluated, and names the class group; (2) the tool automatically generates a unique link accessible to students; (3) each student conducts their evaluation independently and submits the data; (4) the responses are consolidated into a collective dashboard accessible to the instructor. This panel displays an overall average score and averages per framework, a heat map with means and standard deviations per criterion, automatic classification of critical criteria (mean ≤ 1.5), weak criteria (between 1.5 and 2.5), and strong criteria (≥ 3.0), alongside consolidated CSV export.

Table 3

Structure of the AI Analysis Prompt and Pedagogical Objectives

Report Section	Pedagogical Objective
1. Overall Interface Overview	Develop systemic vision: interpret per-framework averages as global quality indicators.
2. Identified Strengths	Reinforce good practices with reference to the corresponding framework.
3. Main Problems and Weak Points	Develop causal reasoning: describe impact on user experience with theoretical grounding.
4. Areas of Divergence (class mode)	Stimulate metacognitive reflection on inter-rater variability.
5. Improvement Recommendations	Develop propositional thinking with concrete, prioritized recommendations.
6. Conclusion	Consolidate synthesis and model technical communication of results.

Source: Research Data.

In accordance with the guidelines proposed by Navarro et al. (2026) for reporting LLM-integrated systems in HCI, UXMind includes an explicit disclaimer regarding the nature of the generated analyses: the report is positioned as an interpretive support instrument, not as a substitute for specialist judgement. Students are encouraged to treat the analysis as a starting point for discussion—a stance consistent with the limitations of LLMs documented in the literature, particularly with regard to the recognition of specific UI conventions and the identification of problems that manifest across multiple screens (Zhong et al., 2025).

Pedagogical Implications

UXMind's design draws on elements of scaffolded learning, collaborative learning, and problem-based learning. The way in which the criteria are presented operates as a cognitive scaffold: for each item, the tool provides the technical definition of the principle being assessed, concrete examples of positive and negative application, and a scoring scale with descriptive labels. This support tends to reduce the extrinsic cognitive load (Sweller, 1988) associated with the need to simultaneously retrieve and apply theoretical concepts and evaluative criteria, freeing the student's attention for the interface inspection process. The generation of AI analyses adds a second layer of scaffolding: by offering a technical

interpretation of the collected data, the system models the expected analytical reasoning—pattern recognition, articulation between evidence and theoretical principles, and formulation of recommendations (Collins et al., 1989).

The Class Mode puts into practice principles of collaborative learning (Dillenbourg, 1999): by confronting their individual evaluations with the class consensus, students encounter divergent perspectives that may give rise to constructive cognitive conflict. The standard deviation displayed on the dashboard makes inter-rater variability visible—a pedagogically relevant finding that empirically illustrates the "evaluator effect" documented by Hertzum and Jacobsen (2001). The collective analysis generated by the LLM deepens this dimension by dedicating a specific section to "Areas of Divergence Among Evaluators", offering students an interpretive model that treats inter-rater variability not as noise, but as information about the complexity of the interface being evaluated.

Considered from a broader perspective, the use of UXMind may contribute to the development of competencies relevant to UX education: by navigating the 47 criteria across four frameworks, interpreting dashboards with consolidated data from multiple evaluators, and critically discussing AI-generated reports, students practise the identification and categorisation of usability problems, the articulation between empirical evidence and theoretical principles, the synthesis of divergent perspectives, and the technical communication of results—competencies that Cockton et al. (2012) associate with the training of UX specialists.

Discussion

The results allow UXMind to be positioned in relation to the state of the art in AI-assisted heuristic evaluation tools. Two characteristics distinguish it from the approaches reported in the literature. The first concerns the division of roles between human and model: whereas Hsueh et al. (2024) and Zhong et al. (2025) propose systems in which the LLM conducts the evaluation autonomously—analysing interfaces directly from screenshots—, UXMind adopts a hybrid approach in which the evaluation is conducted by students and the LLM acts exclusively at the stage of interpreting and synthesising the collected data.

This choice is pedagogically motivated: the formative value of the tool lies precisely in the evaluation process conducted by the student, not in its replacement by AI. The second distinctive characteristic is the integration of four frameworks into a single instrument, whereas the studies by Hsueh et al. (2024) and Zhong et al. (2025) are confined to Nielsen's heuristics. This integration tends to broaden diagnostic coverage and exposes students simultaneously to multiple theoretical perspectives on interface quality—which aligns with the evidence from Cockton and Woolrych (2001) that distinct frameworks capture different categories of problems.

Three limitations of the artefact warrant acknowledgement. The first is the connectivity dependency of the AI analysis component, which restricts its use in contexts with limited internet access. The second is the absence of data persistence across sessions in Class Mode: the browser's LocalStorage retains data only within the same session and device, which prevents access to evaluations from one day to the next without prior export. The third is the inherent variability of LLM outputs: distinct executions of the same prompt may produce analyses with different emphases, limiting the reproducibility of the reports—a characteristic

documented in the literature (Navarro et al., 2026) that reinforces the pertinence of treating the generated analysis as a starting point for discussion rather than a definitive result.

Conclusion

This article presented UXMind, an intelligent agent based on language models conceived to support students in learning interface evaluation in UX/HCI education. Anchored in the Design Science Research methodological approach (Hevner et al., 2004; Peffers et al., 2007), the work proposed and described a technological-pedagogical artefact that brings together 47 evaluation criteria drawn from four consolidated scientific frameworks—Nielsen (1994), Bastien and Scapin (1993), WCAG 2.1 (W3C, 2024), and the Gestalt principles (Wertheimer, 1923, as cited in Lidwell et al., 2003)—, collaborative classroom evaluation dynamics through Class Mode, and LLM-generated interpretive analyses from the data collected during the evaluation process.

The central contribution of this work lies in an articulation that remains underexplored in the literature: the integration of multidimensional heuristic evaluation, pedagogical collaboration, and generative AI in the context of undergraduate Computer Science programmes. UXMind occupies a space identified as underexplored—the application of GenAI to the interface evaluation stage in UX/HCI workflows (Kumar et al., 2025)—with an emphasis on developing the analytical competencies of students in training. The pedagogical implications of the artefact suggest potential to support higher-order competencies in HCI—the identification and categorisation of usability problems, the articulation between evidence and theoretical principles, the synthesis of divergent perspectives, and the technical communication of results—, objectives that traditional instruments such as printed checklists or spreadsheets tend not to address in an integrated manner (Cockton et al., 2012).

Directions for Future Work

The results of this work point to three axes of future investigation. The first is the empirical evaluation of the artefact, through a controlled study comparing the pedagogical efficacy of UXMind with that of traditional methods, measuring the quality of the heuristic evaluations produced, the development of analytical competencies over the course of a semester, and student perceptions of the tool's utility and usability—for instance, through administration of the SUS (System Usability Scale). The second axis involves functional extensions: data persistence via a lightweight backend (such as Firebase or Supabase) to overcome the LocalStorage limitation; integration of multimodal analysis capability, enabling the submission of interface images to the LLM; and development of a longitudinal student progress tracking module. The third axis is expansion to other contexts: UXMind's modular architecture allows for the incorporation of additional frameworks and the adaptation of Class Mode to collaborative professional evaluation settings, such as accessibility audits in public sector organisations.

The growing demand for UX professionals in the post-pandemic market (Verified Market Research, 2025) and the progressive integration of AI into interface design workflows indicate that tools such as UXMind tend to remain relevant in the training of professionals capable of operating at the intersection of heuristic evaluation, collaboration, and artificial intelligence. This work represents an initial step in that direction.

Acknowledgements

To the Samsung Ocean 2.0 Research and Development (R&D) project, from which this article originated, funded by Samsung through resources provided under the Informatics Law for the Western Amazon (Federal Law No. 8,387/1991), and whose dissemination is in compliance with the provisions of Article 39 of Decree No. 10,521/2020.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that ChatGPT 5.3, an AI-based language model, was used in proofreading and refining the language used in the manuscript. The usage was limited to correcting grammatical and spelling errors and rephrasing statements for accuracy and clarity. Additionally, Claude was used as a coding assistant in the development of the UXMind.app. The author further declares that, apart from ChatGPT 5.3 and Claude, no other AI or AI-assisted technologies have been used to generate content in writing the manuscript or developing the software. The ideas, design, procedures, findings, analyses, and discussion are originally written and derived from careful and systematic conduct of the research.

References

Alomari, H. W., Ramasamy, V., Kiper, J. D., & Potvin, G. (2020). A user interface (UI) and user experience (UX) evaluation framework for cyberlearning environments in computer science and software engineering education. *Heliyon*, *6*(5), e03917. https://doi.org/10.1016/j.heliyon.2020.e03917

Bastien, J. M. C., & Scapin, D. L. (1993). Ergonomic criteria for the evaluation of human-computer interfaces. *Technical Report INRIA nº 156*. Institut National de Recherche en Informatique et en Automatique.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Cockton, G., & Woolrych, A. (2001). Understanding inspection methods: Lessons from an assessment of heuristic evaluation. In A. Blandford & J. Vanderdonckt (Eds.), *People and computers XV — Interaction without frontiers* (pp. 171–191). Springer. https://doi.org/10.1007/978-1-4471-0353-0_11

Cockton, G., Woolrych, A., Hornbæk, K., & Frøkjær, E. (2012). Inspection-based evaluations. In J. A. Jacko (Ed.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (3rd ed., pp. 1279–1298). CRC Press. https://www.taylorfrancis.com/chapters/edit/10.1201/b11963-ch-56/inspection-based-evaluations-gilbert-cockton-alan-woolrych-kasper-hornb%C3%A6k-erik-fr%C3%B8kj%C3%A6r

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 1–19). Elsevier. https://telearn.hal.science/hal-00190240

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2023). My AI wants to know if this will be on the exam: Testing OpenAI Codex on CS2 programming exercises. *In Proceedings of the 25th Australasian Computing Education Conference* (pp. 1–10). ACM. https://dl.acm.org/doi/pdf/10.1145/3576123.3576134

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., … Rienties, B. (2025). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, *44*(11), 2518–2544. https://doi.org/10.1080/0144929X.2024.2394886

Henry, S. L., Abou-Zahra, S., & Brewer, J. (2014). The role of accessibility in a universal Web. *In Proceedings of the 11th Web for All Conference* (Article 17, pp. 1–4). ACM. https://dl.acm.org/doi/10.1145/2596695.2596719

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*(4), 421–443. https://doi.org/10.1207/S15327590IJHC1304_05

Hsueh, N.-L., Lin, H.-J., & Lai, L.-C. (2024). Applying Large Language Model to User Experience Testing. *Electronics*, *13*(23), 4633. https://doi.org/10.3390/electronics13234633

Kumar, T., Tu, X., & Zallio, M. (2025). How generative AI is reshaping UI/UX design workflows: A systematic review. *In Human factors in design, engineering, and computing* (Vol. 199, pp. 2438–2452). AHFE International. https://doi.org/10.54941/ahfe1007056

LaBerge, L., O'Toole, C., Schneider, J., & Smaje, K. (2020, October 5). *How COVID-19 has pushed companies over the technology tipping point—and transformed business forever*. McKinsey & Company. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever

Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence,* 9, Article 100445. https://doi.org/10.1016/j.caeai.2025.100445

Lidwell, W., Holden, K., & Butler, J. (2010). *Universal principles of design: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design* (Rev. and updated ed.). Rockport Publishers.

Nassar, V. (2012). Common criteria for usability review. *Work: A Journal of Prevention, Assessment & Rehabilitation*, v. 41, (supl. 1), 1053–1057. https://journals.sagepub.com/doi/10.3233/WOR-2012-0282-1053

Navarro, K. F., Syriani, E., & Arawjo, I. (2026). *Reporting and reviewing LLM-integrated systems in HCI: Challenges and considerations* [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2602.05128

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 152–158). ACM. https://doi.org/10.1145/191666.191729

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *In Proceedings of the ACM CHI'90 Conference on Human Factors in Computing Systems* (pp. 249–256). ACM. https://doi.org/10.1145/97243.97281

Ntoa, S. (2025). Usability and User Experience Evaluation in Intelligent Environments: A Review and Reappraisal, *International Journal of Human–Computer Interaction*, *41*(5), 2829–2858. https://doi.org/10.1080/10447318.2024.2394724

Peffers K., Tuunanen T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research, *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., & Santos, E. A. (2024). "It's weird that it knows what I want": Usability and interactions with Copilot for novice programmers. *ACM Transactions on Computer-Human Interaction*, *31*(1), Article 4. https://doi.org/10.1145/3617367

Soto-Acosta, P. (2020). COVID-19 pandemic: Shifting digital transformation to a high-speed gear. *Information Systems Management*, *37*(4), 260–266. https://doi.org/10.1080/10580530.2020.1814461

UXness Survey. (2024). *2024 Annual UX Tools Survey*. https://www.uxness.in/2024/06/2024-annual-ux-tools-survey-by-uxness.html

Verified Market Research. (2025). User experience (UX) market size, share & forecast. https://www.verifiedmarketresearch.com/product/user-experience-ux-market/

W3C. (2024). Web content accessibility guidelines (WCAG) 2.1. *W3C Recommendation*, 05 December 2024. https://www.w3.org/TR/WCAG22/

Zhong, R., McDonald, D. W., & Hsieh, G. (2025). *Synthetic heuristic evaluation: A comparison between AI- and human-powered usability evaluation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.02306

Contact email: dnoronha@uea.edu.br