CONFERE'26
2 ET 3 JUILLET 2026, GZIRA, MALTE.

A SYSTEMATIC DIAGNOSTIC TOOL FOR LLM INTEGRATION IN SERIOUS GAMES: BRIDGING PEDAGOGICAL NEEDS AND AI FUNCTIONAL ROLES IN PRODUCT ENGINEERING DESIGN

Xinyu LIU [1 2], Frédéric SEGONDS [1], Ruding LOU [2], Camille JEAN [1]
[1] Laboratoire de Conception de Produits et Innovation LCPI - Arts et Métiers Institute of Technology - 151 Boulevard de l'Hôpital, 75013 Paris, France

[2] Laboratoire d'Ingénierie des Systèmes Physiques et Numériques LISPEN - Arts et Métiers Institute of Technology - 44 quai Saint Cosme, 71100 Chalon-sur-Saône, France

Abstract : Integrating AI into educational games for Product Engineering Design presents notable opportunities, yet developers and educators often face challenges in aligning AI functionalities with specific pedagogical goals. To explore a structured approach, this paper applies an AI function taxonomy to develop AI Architecture Audit System. The tool categorizes AI into four primary roles: Social (supporting teamwork), Tutors (offering adaptive guidance), Assessors (evaluating design choices), and Generators (creating relevant scenarios). By mapping pedagogical needs to potential technical solutions, the tool aims to reduce early-stage design ambiguity and encourage purpose-driven AI integration. A pilot usability study with engineering and industrial design graduate students suggests that the tool helps clarify decision-making and translates preliminary concepts into technical specifications. This work offers a baseline framework for AI-assisted serious game design, with further classroom testing needed to verify and refine its broader applicability.

Keyword : Product Design, Engineering Education, Artificial Intelligence Agent, Serious Games.

1 INTRODUCTION

Product Engineering Design Education increasingly fosters systemic thinking and multidisciplinary collaboration through Serious Games [1]. However, generative AI integration faces assessment accuracy limitations [2] and lacks explicit pedagogical guidance, often marginalizing educators [3] and risking underperforming "solution-driven" biases. Consequently, a systematic methodology to map learning barriers to AI roles remains absent and balancing industrial and pedagogical design early on remains challenging [4]. To address this, this study develops the AI Architecture Audit System, an automated, interactive web tool designed to bridge the pedagogical-technical divide. Grounded in a structured ex-ante heuristic framework, the system translates unstructured pedagogical pain points identified by educators into formalized AI functional roles and quantifiable engineering specifications. By automating early-stage design alignment, this tool aims to reduce design ambiguity, mitigate solution-driven biases, and provide scalable, real-time guidance for multidisciplinary development teams before high-cost prototyping begins.

2 STATE OF ART

Contemporary engineering education increasingly leverages serious games, yet the inclusion of generative AI reveals a critical mismatch: existing design heuristics rely heavily on generic usability metrics [5] and completed prototypes [6, 7] rather than addressing the unique traits of interactive

LLMs [8]. Resolving this tension requires a fundamental understanding of how AI agents function as autonomous educational entities rather than mere script triggers. Consequently, serious game design must map specific learning barriers to clear taxonomic dimensions—namely social, tutoring, assessing, and generating functions. Aligning these qualitative pedagogical goals with a quantitative capability matrix to measure AI agency provides the technical blueprint necessary to move away from high-cost, trial-and-error development toward predictive, ex-ante design diagnostics.

2.1. AI Agents in serious game

While LLM-driven AI agents have transitioned from script triggers to autonomous educational entities [8], with documented potential in automating tasks [9] and delivering scaffolding feedback [10], a systematic framework defining their functional boundaries is still lacking, leaving evaluations confined to generic usability [5]. This study addresses this limitation by deploying a Functional Taxonomy (Table 1) that refines AI roles into four core dimensions: Social, Tutors, Assessors, and Generators. This framework integrates traditional Intelligent Tutoring Systems (ITS) with generative advancements like emotion recognition, dynamic difficulty adjustment, and NLP-driven narratives using reusable components [11]. Ultimately, by explicitly mapping pedagogical objectives to each role, the taxonomy provides a theoretical anchor for serious game design diagnostics [12].

Table 1. Taxonomy of core AI functions.

AI Core Functions	Functional Definition	Core Pedagogical Goal
Social	Facilitates team collaboration and exchange.	Fosters socialized learning and peer exchange.
Tutors	Monitors progress and provides adaptive feedback.	Delivers real-time, adaptive instructional scaffolding.
Assessors	Logs behaviors and evaluates engineering decisions.	Formatively evaluates learners' design-thinking choices.
Generators	Uses LLMs to dynamically generate designs.	Contextualizes dynamic and authentic engineering scenarios.

2.2. Limits of Existing Evaluation Scales: The Need for Design Diagnosis

While mature evaluation scales exist for serious games, they exhibit distinct limitations when integrating LLMs. First, most methods rely on "ex-post evaluation." Although heuristic evaluation uncovers interface usability issues [6], it requires existing prototypes or systems [7]. For LLM-driven games, designers require pre-development guidance on AI maturity and functional pathways, which existing scales fail to provide [13]. Second, a domain mismatch exists most educational game heuristics still rely on generic usability guidelines [5], lacking specificity regarding gamified motivational stimulation and flow maintenance [12]. Finally, traditional scales fail to cover criteria unique to conversational agents, such as context preservation, privacy, and content authenticity [8]. This lag disconnects pedagogical needs from technical implementation—a gap this study fills via an ex-ante evaluation framework.

2.3. Measuring AI Agency: From Autonomy to Maturity

To quantify AI involvement in pedagogical interventions, an AI Functional Capability Scoring Matrix was developed. Referencing Levels of Automation theory, it categorizes AI intervention into maturity levels 0 to 3, as detailed in Table 2:

Table 2. AI functional capability scoring matrix.

Type	0: None	1: Basic	2: Interactive	3: Specialized
Social	No interaction	One-way command response	Basic multi-turn dialogue	Distinct persona & context awareness
Tutors	No guidance	Static info presentation	Basic task instructions	Dynamic difficulty & path adaptation
Assessors	No feedback	Simple right/wrong binary	Rule-based auto-scoring	Qualitative diagnosis & logic tracing

Generators	No generation	Template-based retrieval	Assisted content tweaking	Fully original content creation

Although these scores are internalized as specific heuristic questions within the automated tool, their core logic remains grounded in this maturity model, aiming to guide designers in determining whether AI exists as an "auxiliary tool" or a "core engine." This stepped logic from perception to complex decision-making aligns with the classification principles of automated system functions in human-computer interaction [14]. Even though the heuristic tool employs descriptive diagnostic language, the underlying framework is still supported by the logic of technical maturity levels 0–3. Concretely, this 0–3 capability matrix establishes a foundational rubric for the inference engine, allowing the AI's qualitative maturity levels to be systematically translated and integrated into the multi-criteria continuous quantitative metrics of the web platform.

2.4. The Paradigm Shift: Pre-guidance Heuristics in Design

Most design methodologies rely on late-stage evaluation, but a "develop-first, test-later" model in serious games yields exorbitant rework costs due to technical and pedagogical complexities [15]. Traditional heuristics act primarily as ex-post inspections, failing to provide early directional guidance [6]. Responding to the paradigm shift toward ex-ante guidance, this study leverages functional superposition theory [11] to translate AI role collaborative logic into an actionable diagnostic tool. This bridges the cognitive gap between educators and engineers [9], ensuring AI integration aligns with pedagogical needs at the outset to prevent later-stage logic mismatches and system failures [5].

3 SYSTEM ARCHITECTURE

Focusing on core diagnostics, the AI Architecture Audit System operationalizes the ex-ante heuristic framework into an automated inference engine that deconstructs unstructured user input, aligns it with the AI taxonomy, and computes semantic fit metrics.

3.1. The 4-Step Decision Path

The AI Architecture Audit System operationalizes a structured workflow to precisely translate vague pedagogical ideas into concrete technical requirements. Serving as an automated implementation of the ex-ante heuristic framework, this system converts qualitative design criteria into an LLM-driven inference engine. Grounded in these metrics, the system architecture (Figure 1) comprise three core processing zones spanning four evolutionary phases.

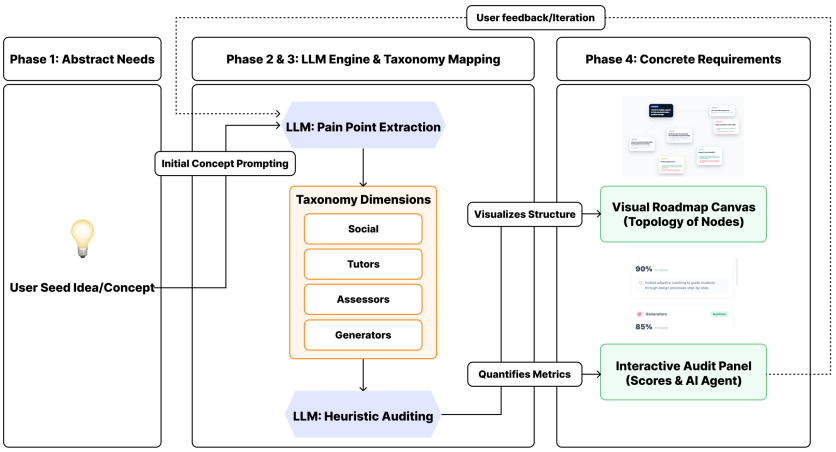

Figure 1. System architecture and workflow.

Initiated by a User Seed Idea, the system deconstructs unstructured input via the LLM: Pain Point Extraction module into explicit pedagogical pain points. Mapped to the core taxonomy dimensions (Social, Tutors, Assessors, Generators), the LLM: Heuristic Auditing module executes quantitative and qualitative evaluations following the linear pathway illustrated in Figure 2. Finally, this processing translates into interactive design outcomes: the Visual Roadmap Canvas presents structural logic as

topological nodes, while the Interactive Audit Panel provides Fit Scores for metric quantification. This architecture establishes a closed-loop feedback system for continuous User Feedback/Iteration.

3.2. LLM-based Heuristic Auditing

As a heuristic inference engine, the LLM bridges the pedagogical-technical gap by aligning the unstructured User Seed Idea Concept with a product engineering knowledge base through semantic enhancement. Following the linear pathway in Figure 2, this conversion executes in four sequential phases: (1) AI Analysis Semantic extracts explicit pedagogical Pain Points (A/B) via semantic analysis; (2) Multidimensional AI Function Mapping channels these pain points into four distinct dimensions (AI Functions A–D) to prevent single-solution bias; (3) AI Expansion (User Click) triggers interactive context expansion to retrieve design patterns and case studies; and (4) Detailed Plan integrates these insights to output alternative Detailed Plans (A–C) as an executable technical blueprint.

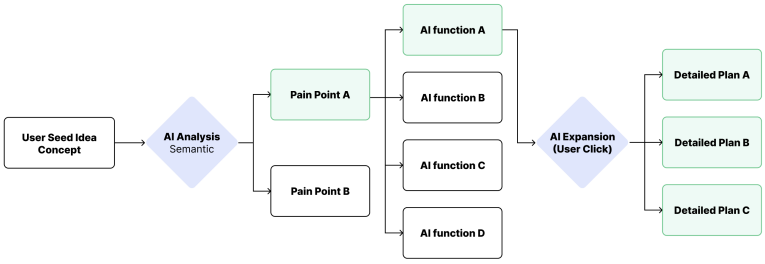

Figure 2. LLM: Pain Point Extraction.

Through this systematic generation pipeline, the tool transforms traditional, experience-based trial-and-error into an automated evaluation process driven by semantic alignment and logical deduction. This drastically enhances the scientific rigor of preliminary Serious Games design while ensuring the precise positioning of AI technology in pedagogical applications.

3.3. Mechanism of Semantic Alignment Scoring

To transform qualitative design visions into actionable metrics, the system translates the discrete capability levels (0–3) from Table 2 into a continuous 100-point Fit Score via a multi-criteria Weighted Sum Model (WSM) [16, 17]. The LLM engine first assesses the input concept against the taxonomy rubrics, utilizing the 0–3 maturity matrix to benchmark individual scores for each AI function. These dimensions are then mathematically aggregated using a "4:4:2" weighting ratio aligned with classic product engineering design principles [18, 19]: Task Relevance (40%, mapped from the AI's capability alignment), Pain-point Coverage (40%, evaluating objective mitigation), and Implementation Feasibility (20%, bounded by the required maturity level constraints).

Within this framework (Figure 3), the system invokes the DeepSeek API for automated scoring via preset rubrics. While advanced models acting as evaluators (LLM-as-a-judge) align well with human preferences [20], using an LLM to assess how other LLMs should be integrated introduces a risk of circular bias, where the evaluator inherently favors its own generative logic. To mitigate this self-referential loop, our system decouples evaluation by introducing independent 0–3 maturity rubrics and exogenous constraints from classic engineering design principles [14, 19]. This structuring ensures that the early-stage scoring remains highly structured, interpretable, and reproducible. While the semantic mapping to the 0–3 rubrics is still parsed by the LLM, the subsequent aggregation relies on static deterministic mathematical formulations (WSM), ensuring that the final Fit Score is not a direct product of unconstrained generative heuristics.

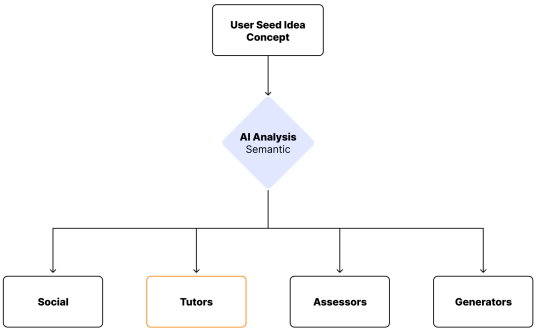

Figure 3. Semantic analysis and dimensional mapping.

4 SYSTEM DESIGN AND IMPLEMENTATION

Moving from the underlying processing logic to practical implementation, the focus shifts to the user interface and interactive layout of the developed tool. The theoretical mechanisms described previously are here operationalized into a functional design workspace to support intuitive human-AI collaboration.

4.1. Conceptual Input Module

The AI Architecture Audit System serves as the core entry point of the system to externalize preliminary conceptions (Figure 4). Adhering to minimalist design principles, the interface features only a low-barrier interactive text box. This design significantly reduces the user's cognitive load, allowing instructors or developers to directly input vague, unstructured pedagogical concepts to provide the raw driving force for the subsequent generation workflow.

Figure 4. User interface of the input module.

The system utilizes a multi-model integrated backend architecture that flexibly invokes different mainstream LLMs based on task requirements, currently relying on the DeepSeek (deepseek-chat) API due to its exceptional reasoning capabilities and high cost-effectiveness. Upon submission, the system leverages the model's semantic understanding to perform intent recognition and feature extraction on unstructured input, precisely capturing pedagogical intent. Utilizing a preset prompt strategy, the system transforms the vague Seed Idea into a structured Semantic Seed to establish the logical foundation for automated generation in subsequent nodes. Crucially, the architecture maintains robust scalability, allowing the subsequent integration of other high-performance models like GPT-4 for complex, cross-domain demands.

4.2. Dynamic Canvas & Logic Visualization

The system's core interaction area features a React-based Infinite Canvas (Figure 5) that transforms abstract pedagogical logic into a topological Node Network. Within this non-linear canvas, the logical evolution is structured through three node layers: the centered Core Theme Node holding the initial user concept, the middle Pain Point Nodes anchoring practical instructional issues, and the outer AI Function Nodes classified into four core roles (Social, Tutors, Assessors, and Generators) to deliver concrete implementation strategies. Through connecting lines, the canvas demonstrates a problem-solving design logic, using the lucide-react library for semantic icon annotation to enhance readability. Additionally, smooth zoom and pan operations ensure efficient navigation, providing a fluid collaborative environment within large-scale networks.

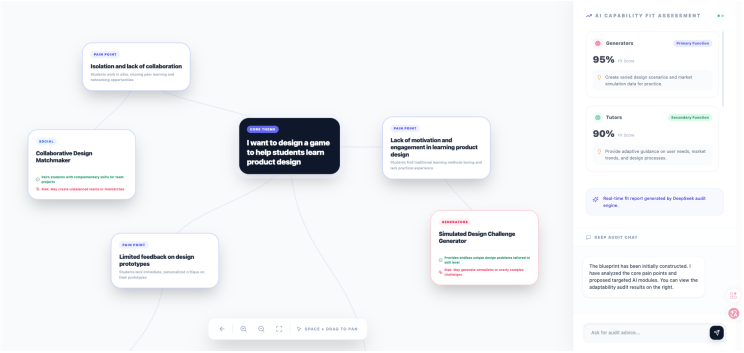

Figure 5. Interactive node network on the dynamic canvas.

4.3. Real-time Audit & Decision Support

An automated background Audit Engine triggers real-time updates (updateRecommendations) to ensure design consistency. The Fit Assessment module scores and labels the four AI roles (Social, Tutors, Assessors, Generators) as "Primary" or "Secondary" for prioritization, while a Deep Audit Chat allows node modifications to enable transparent co-design. Scientifically, the Semantic Alignment Score evaluates similarity between pedagogical objectives and AI strategies. This Evidence-based Design approach effectively mitigates LLM hallucinations, ensuring rigor and reliability.

5 RESULTS: PILOT STUDY AND USER FEEDBACK

A pilot usability study was conducted with two graduate students from industrial design and engineering backgrounds, focusing on how the tool guides students in clarifying design directions and fostering cross-disciplinary consensus. Results show the system significantly enhanced decision-making transparency and precision. It guided participants to identify specific Pain Point nodes from vague concepts and leverage quantified Fit Scores to determine the Assessors role as a Primary Function for complex decision-making.

Qualitative feedback from the pilot study suggests that the tool has the potential to bridge the gap between pedagogical intent and technical implementation. First, regarding decision-making transparency, the participants indicated that the Semantic Alignment Score provided a structured reference that helped transition design discussions from intuition-driven to more logic-driven rationales. Second, the tool appeared to facilitate cross-disciplinary communication by generating actionable pre-plans. Specifically, the engineering-background participant noted that the Deep Audit Chat delivered heuristic functional concepts rather than rigid specifications, translating vague requirements into a collaborative framework for rapid alignment. Ultimately, within the scope of this pilot test, the tool showed promise as a common language to narrow cognitive gaps in cross-disciplinary collaboration, helping to expand the exploration space for design concepts and guide preliminary thinking divergence based on a structured heuristic framework.

6 CONCLUSION AND FUTURE WORK

This study refines the AI function taxonomy by enabling roles to compound flexibly to meet multi-layered pedagogical demands. Empirical results from the AI Architecture Audit System prove its viability as a diagnostic tool. By visualizing and quantifying these configurations ex-ante, the system eliminates early-stage ambiguity, prevents late-stage rework, and helps cross-disciplinary teams optimize AI integration.

Future interdisciplinary research will leverage empirical pedagogical pain points in smartly connected product design (e.g., hardware-software co-design) as bottom-up inputs to iterate the tool's LLM knowledge base. Ultimately, guided by the system's metrics, several AI-enhanced Serious Game

prototypes will be classroom-tested, establishing a closed-loop validation framework—where pain points drive design, tools empower development, and products feedback into theory—across the design lifecycle.

7 REFERENCES

[1] PACHECO-VELÁZQUEZ, Ernesto, David Ernesto SALINAS-NAVARRO a María Soledad RAMÍREZ-MONTOYA. Serious Games and Experiential Learning: Options for Engineering Education. *International Journal of Serious Games* [online]. 2023, **10**(3), 3–21. ISSN 2384-8766. Dostupné z: doi:10.17083/ijsg.v10i3.593

[2] NIKOLIC, Sasha, Scott DANIEL, Rezwanul HAQUE, Marina BELKINA, Ghulam M. HASSAN, Sarah GRUNDY, Sarah LYDEN, Peter NEAL a Caz SANDISON. ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education* [online]. 2023, **48**(4), 559–614. ISSN 0304-3797. Dostupné z: doi:10.1080/03043797.2023.2213169

[3] ZAWACKI-RICHTER O., V. I. MARÍN, M. BOND a F. GOUVERNEUR. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education* [online]. 2019, **16**(1), 39. ISSN 2365-9440. Dostupné z: doi:10.1186/s41239-019-0171-0

[4] TOMIYAMA, Tetsuo, P. GU, Yan JIN, Eric LUTTERS, Christian KIND a Fumihiko KIMURA. Design methodologies: Industrial and educational applications. *CIRP Annals-Manufacturing Technology* [online]. 2009, **58**, 543–565. Dostupné z: doi:10.1016/j.cirp.2009.09.003

[5] Heuristic Evaluation on Usability of Educational Games: A Systematic Review. *Informatics in Education - An International Journal*. 2019, **18**(2), 427–442. ISSN 1648-5831.

[6] NIELSEN, Jakob. Finding usability problems through heuristic evaluation. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* [online]. New York, NY, USA: Association for Computing Machinery, 1992, s. 373–380 [vid. 2026-06-08]. CHI '92. ISBN 978-0-89791-513-7. Dostupné z: doi:10.1145/142750.142834

[7] NIELSEN, Jakob a Rolf MOLICH. Heuristic evaluation of user interfaces. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* [online]. New York, NY, USA: Association for Computing Machinery, 1990, s. 249–256 [vid. 2026-06-08]. CHI '90. ISBN 978-0-201-50932-8. Dostupné z: doi:10.1145/97243.97281

[8] LANGEVIN, Raina, Ross J LORDON, Thi AVRAHAMI, Benjamin R. COWAN, Tad HIRSCH a Gary HSIEH. Heuristic Evaluation of Conversational Agents. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* [online]. New York, NY, USA: Association for Computing Machinery, 2021, s. 1–15 [vid. 2026-06-08]. CHI '21. ISBN 978-1-4503-8096-6. Dostupné z: doi:10.1145/3411764.3445312

[9] CHU, Zhendong, Shen WANG, Jian XIE, Tinghui ZHU, Yibo YAN, Jinheng YE, Aoxiao ZHONG, Xuming HU, Jing LIANG, Philip S. YU a Qingsong WEN. *LLM Agents for Education: Advances and Applications* [online]. B.m.: arXiv. 4. únor 2026 [vid. 2026-06-08]. Dostupné z: doi:10.48550/arXiv.2503.11733. arXiv:2503.11733 [cs.CY]

[10] CHIEN, Chih-Chung, Hung-Yu CHAN a Huei-Tse HOU. Learning by playing with generative AI: design and evaluation of a role-playing educational game with generative AI as scaffolding for instant feedback interaction. *Journal of Research on Technology in Education* [online]. 2025, **57**(4), 894–913. ISSN 1539-1523. Dostupné z: doi:10.1080/15391523.2024.2338085

[11] WESTERA, Wim, Rui PRADA, Samuel MASCARENHAS, Pedro A. SANTOS, João DIAS, Manuel GUIMARÃES, Konstantinos GEORGIADIS, Enkhbold NYAMSUREN, Kiavash BAHREINI, Zerrin YUMAK, Chris CHRISTYOWIDIASMORO, Mihai DASCALU, Gabriel GUTU-ROBU a Stefan RUSETI. Artificial intelligence moving serious gaming: Presenting reusable game AI components. *Education and Information Technologies* [online]. 2020, **25**(1), 351–380. ISSN 1573-7608. Dostupné z: doi:10.1007/s10639-019-09968-2

[12] TONDELLO, Gustavo F., Dennis L. KAPPEN, Elisa D. MEKLER, Marim GANABA a Lennart E. NACKE. Heuristic Evaluation for Gameful Design. In: *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* [online]. New York, NY, USA: Association for Computing Machinery, 2016, s. 315–323 [vid. 2026-06-08]. CHI PLAY Companion '16. ISBN 978-1-4503-4458-6. Dostupné z: doi:10.1145/2968120.2987729

[13] LAW, Effie Lai-Chong a Ebba Thora HVANNBERG. Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In: *Proceedings of the third Nordic conference on Human-computer interaction* [online]. New York, NY, USA: Association for Computing Machinery, 2004, s. 241–250 [vid. 2026-06-08]. NordiCHI '04. ISBN 978-1-58113-857-3. Dostupné z: doi:10.1145/1028014.1028051

[14] PARASURAMAN, Raja, Thomas SHERIDAN a Christopher WICKENS. A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 30(3), 286-297. *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society* [online]. 2000, **30**, 286–97. Dostupné z: doi:10.1109/3468.844354

[15] LAW, Effie Lai-Chong a Ebba Thora HVANNBERG. Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In: *Proceedings of the third Nordic conference on Human-computer interaction* [online]. New York, NY, USA: Association for Computing Machinery, 2004, s. 241–250 [vid. 2026-06-08]. NordiCHI '04. ISBN 978-1-58113-857-3. Dostupné z: doi:10.1145/1028014.1028051

[16] TAHERDOOST, Hamed a Mitra MADANCHIAN. Multi-Criteria Decision Making (MCDM) Methods and Concepts. *Encyclopedia* [online]. 2023, **3**(1), 77–87. ISSN 2673-8392. Dostupné z: doi:10.3390/encyclopedia3010006

[17] TRIANTAPHYLLOU, Evangelos. Multi-Criteria Decision Making Methods. In: *Multi-criteria Decision Making Methods: A Comparative Study* [online]. B.m.: Springer, Boston, MA, 2000 [vid. 2026-06-08], s. 5–21. ISBN 978-1-4757-3157-6. Dostupné z: doi:10.1007/978-1-4757-3157-6_2

[18] PAHL, G. *Engineering design : a systematic approach /*. 1988.

[19] ULRICH, Karl a Steven EPPINGER. *EBOOK: Product Design and Development*. B.m.: McGraw Hill, 2011. ISBN 978-0-07-714396-1.

[20] ZHENG, Lianmin, Wei-Lin CHIANG, Ying SHENG, Siyuan ZHUANG, Zhanghao WU, Yonghao ZHUANG, Zi LIN, Zhuohan LI, Dacheng LI, Eric XING, Hao ZHANG, Joseph GONZALEZ a Ion STOICA. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*. 2023, **36**, 46595–46623.

Main contact : Xinyu LIU

Contact information : xinyu.liu@ensam.eu