AI in Usability Testing

University of Turku
Master of Science in Technology Thesis
Department of Computing
Software Engineering
June 2025
Jannatul Tazrin Biva

Supervisor:
Prof. Jaakko Järvi

UNIVERSITY OF TURKU
Department of Computing

Jannatul Tazrin Biva: AI in Usability Testing

Master of Science in Technology Thesis, Software Engineering, 52 p.
June 2025

This thesis looks at how artificial intelligence can be used in usability testing. A custom API (Application Programming Interface) was built that works with GPT-4o to examine screenshots of user interfaces and check for usability problems. The tool was tested on four online unit converter websites. Its findings were then compared with feedback from students who reviewed the same websites as part of a class project. The study shows that AI can efficiently and consistently detect usability problems, especially in areas such as visibility, error prevention, and system feedback. At the same time, it highlights the value of human input. Students were able to offer personal insights and noticed details specific to each context that AI alone could not capture. The findings suggest that a hybrid approach, combining automated analysis with human judgment, offers a more complete usability testing framework. Future research should aim to improve AI transparency, address potential biases, and refine these tools for wider use in usability evaluation.

Keywords: AI usability testing, Nielsen heuristics, GPT-4o, API, hybrid usability analysis, usability testing automation, AI-human comparison, user experience (UX), GUI snapshot analysis

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Jaakko Järvi, for his continuous support, valuable feedback, and encouragement throughout this thesis process. I am also thankful to Senior Researcher Seppo Helle for kindly providing the student assignments used in this study, which served as essential data for evaluation.

My heartfelt thanks go to my husband for his unwavering support and to my son, whose joy and curiosity kept me motivated. I am especially grateful for the strength and inspiration brought by the arrival of our newborn during this academic journey.

Additionally, I used ChatGPT to support grammar correction and improve the structure of the thesis text, especially during times when balancing studies and family life was most demanding.

Finally, I appreciate the resources and academic environment provided by the University of Turku, which enabled me to complete this research.

Contents

Acknowledgements i

1 Introduction **1**

2 Literature Review **3**

2.1 Traditional Usability Testing and Its Limitations 3

2.2 Transition to AI-Supported Usability Testing 4

2.3 Key AI Techniques in Usability Testing 5

2.3.1 Machine Learning (ML) 6

2.3.2 Natural Language Processing (NLP) 6

2.3.3 Computer Vision 7

2.4 Tools, Frameworks, and Real-Time Capabilities in AI-Driven Usability Testing 8

2.4.1 Real-Time Data Collection and Analysis 8

2.4.2 Scalability and Predictive Analytics 9

2.4.3 AIMT-UXT and AI-driven Tools 9

2.5 Case Studies and Applications 10

2.5.1 Intelligent Medical Systems 10

2.5.2 AI-Driven Design Heuristics in UX Development 11

2.5.3 Smart Online Shopping Systems 11

2.5.4 AI in Various Industries 12

2.6 Recent Advancements in AI for Usability Testing 13

2.6.1 Generative AI and Multi-modal Models 13

2.6.2 Intelligent Automation and Digital Twins 14

2.7 Model Transparency and Ethical Considerations in AI-Driven Usability Testing . . . 15
2.8 Comparison of AI and Human Usability Testing . . . 17
2.8.1 Complementary Strengths . . . 18
2.8.2 Human-AI Integration in Practice . . . 18
2.9 Synthesis and Identified Gaps . . . 19

3 Materials and methods **21**
3.1 Experimental design . . . 21
3.1.1 Participant selection and task design . . . 21
3.1.2 Exclusion of omni calculator . . . 22
3.1.3 Data collection and privacy . . . 22
3.1.4 Comparison and analysis process . . . 22
3.2 API design and implementation . . . 23
3.2.1 Image upload and conversion . . . 23
3.2.2 API call to OpenAI . . . 23
3.2.3 Error handling and response management . . . 24
3.2.4 Analysis and response delivery . . . 24
3.2.5 Follow-up questions . . . 25
3.3 Data analysis . . . 25
3.3.1 Clarification on Severity Ratings . . . 26
3.3.2 Website 1 analysis: Unit Converter . . . 26
3.3.3 Website 2 analysis: Engineering toolbox . . . 30
3.3.4 Website 3 analysis: Convert me . . . 33
3.3.5 Website 4 analysis: Asknumbers . . . 37
3.4 Comparative analysis across all websites . . . 40
3.4.1 Tool ranking by AI and students . . . 41
3.4.2 Usability patterns across heuristics . . . 42

4 Discussion **46**
4.1 AI in usability testing: strengths and limitations . . . 46

4.2 Comparative analysis of AI and human evaluations 47
4.3 Methodological insights and query optimization 48
4.4 Scope and applicability of AI in usability testing 48
4.5 Scalability and practical applications 49
4.6 Future Directions and Enhancements 49

5 Conclusion **51**

References **53**

List of Figures

3.1 Interface of the UnitConverters.net website (screenshot by the author). . . 28

3.2 Comparison of Heuristic Evaluations for UnitConverters.net: AI vs. Students. 30

3.3 Interface of the EngineeringToolbox website (screenshot by the author). . . 31

3.4 Comparison of Heuristic Evaluations for Engineering Toolbox: AI vs. Students 34

3.5 Interface of the Convert me.com website (screenshot by the author). 35

3.6 Comparison of Evaluations for Convert-Me: AI vs. Students. 37

3.7 Interface of the Asknumbers.com website (screenshot by the author). . . . 38

3.8 Comparison of Evaluations for AskNumbers: AI vs. Students. 41

3.9 Students' Rankings of Best and Worst Tools. 42

3.10 Heatmap Comparing AI and Student Severity Ratings Across Heuristics. . 43

List of Tables

3.1 Comparison of average severity ratings between AI and students evaluations for the *UnitConverters.net* website. 30
3.2 Comparison of average severity ratings between AI and students evaluations for the *EngineeringToolbox.com* website. 33
3.3 Comparison of average severity ratings between AI and students evaluations for the *Convert-me.com* website. 37
3.4 Comparison of average severity ratings between AI and students evaluations for the *Asknumbers.com* website. 40
3.5 Comparison of usability issues identified by AI and students. 44

List of acronyms

AI Artificial Intelligence

AIMT-UXT Artificial Intelligence and Mouse Tracking-based User Experience Tool

API Application Programming Interface

DNCF Deep Neural Collaborative Filtering

GUI Graphical User Interface

HCI Human-Computer Interaction

HVAC Heating, Ventilation, and Air Conditioning

IDE Integrated development environment

LSTM Long short-term memory

LUIM Large user interface model

ML Machine learning

NLP Natural Language Processing

UX User Experience

XAI Explainable AI

1 Introduction

Artificial intelligence (AI) has seen rapid growth, transforming many areas of software engineering. In fields like code generation, test automation, and integrated development environments (IDEs), AI-powered tools, often called "co-pilots," assist developers with tasks such as generating code, predicting errors, and even fixing bugs. This shift has brought a lot of improvements in productivity, speed, and accuracy for developers. Nevertheless, despite AI's progress, it has not yet become a regular tool in user experience (UX) design.

Usability evaluation is a vital part of UX, traditionally carried out by human evaluators who observe user interactions, assess usability issues through established guidelines, and analyze feedback with their own judgment. Although human-led evaluations provide valuable insights, they are time-consuming and labor-intensive. This makes it challenging to scale the process, particularly in fast-moving development environments where quick changes are often necessary. AI has the potential to tackle some of these challenges by speeding up the testing process and providing automated, data-driven insights. However, AI in this area is still evolving, and it faces challenges in interpreting subjective, context-based elements—like understanding user emotions, cultural context, and individual preferences—that are essential for a full UX evaluation.

This thesis aims to explore AI's potential role in usability testing, especially looking at how AI could support or even replace human evaluators for specific tasks. The study focuses on a custom-built API that was developed using GPT-4o. This API performs automated heuristic evaluations on online unit converter websites, analyzing visual content to detect usability issues. It specifically looks for problems related to important usability areas such as system feedback, visibility, error prevention, and user control. To test how effective the API is, its results were compared with those provided by student evaluators

who used traditional, human-based evaluation methods.

By conducting this analysis, the study aims to find out some key questions: Can AI replace human usability evaluators? What are the key strategies for effectively incorporating AI into usability testing? And how can AI complement human abilities to create a more efficient and accurate usability testing process? The findings suggest that the API does offer reliable, quick, and objective evaluations, especially in areas like interface visibility, response speed, and structural consistency. However, it falls short in areas that require more subjective judgment, like understanding emotional responses, contextual relevance, and the subtleties of user experience. These results indicate that AI cannot fully replace human evaluators but can work alongside them in a complementary way. This creates a hybrid model where AI takes care of many technical aspects, while human insights address the subjective and experiential layers needed for thorough UX analysis.

In conclusion, this thesis proposes a mixed approach to usability testing. In this model, AI and human evaluators work together within a framework that is both scalable and efficient. By combining AI's strengths in speed and consistency with human evaluators' ability to provide contextual insights, this approach has the potential to make usability testing both more efficient and more effective.

This research adds to ongoing discussions about AI's role in UX design, and shedding light on possible future advancements for AI in usability testing. It also addresses some ethical issues that arise with integrating AI in human-centered fields.

2 Literature Review

2.1 Traditional Usability Testing and Its Limitations

Usability testing is essential for assessing how effectively users can navigate and interact with digital systems. Traditionally, this process involves monitoring participants as they use a product in either a lab-based setting or remotely, aiming to identify areas of confusion or inefficiency. The goal is to identify obstacles, gather feedback, and improve the system's design. Core evaluation metrics commonly include task success rate, completion time, error frequency, and subjective user satisfaction [1].

Despite its strengths, traditional usability testing is often time-consuming, labor-intensive, and costly. It requires careful planning, participant recruitment, structured facilitation, and detailed analysis. Barnum (2011) emphasizes that these processes demand substantial expertise, making traditional testing resource-heavy and less suitable for iterative design or fast-paced development cycles [2]. To mitigate some of these challenges, Jakob Nielsen's "discount usability testing" approach proposes using smaller sample sizes to reduce costs and preparation time while still yielding useful insights [3].

However, even simplified versions of traditional testing face scalability issues. Krug (2010) and Rubin & Chisnell (2008) observe that repeated usability tests during multiple development phases can overwhelm teams, particularly those with limited time or personnel [1], [4]. In addition, traditional testing methods usually deliver insights only after sessions have concluded, which may delay the integration of feedback into the design process [5].

These limitations are becoming increasingly incompatible with agile and continuous deployment practices, where timely and ongoing evaluation is essential. As a result, researchers have begun to explore AI-based alternatives that promise more scalable, efficient,

and real-time usability testing solutions.

2.2 Transition to AI-Supported Usability Testing

To address the limitations of traditional usability testing—namely high costs, limited scalability, and delayed results—researchers have turned to artificial intelligence (AI) to automate and enhance the process. AI-powered tools aim to deliver faster, scalable insights while reducing the logistical demands of conventional approaches.

At the heart of this transformation is the integration of AI technologies—including machine learning, natural language processing, and computer vision—into multiple phases of usability evaluation. These tools enable the automated detection of user interaction patterns, emotional cues, and textual feedback—elements that were previously gathered through manual observation or post-session surveys [6]–[9].

Among these technologies, machine learning has been particularly effective in identifying hidden usability issues. By analyzing large volumes of user interaction data, ML models can uncover recurring patterns, predict areas where users may struggle, and group users based on behavioral similarities. These insights support early-stage design changes and promote ongoing improvements, especially within agile development environments [6], [7].

In contrast, natural language processing focuses on analyzing textual feedback in real time. By examining user comments from surveys, support chats, and online reviews, NLP tools can surface key themes and sentiment trends that might otherwise go unnoticed, offering a deeper understanding of user experiences [8].

Computer vision technologies further extend usability analysis by tracking visual and emotional engagement.In both academic and commercial settings, techniques like eye-tracking, facial expression monitoring, and gesture recognition are being increasingly applied to study user interactions with digital interfaces. These insights help designers understand which elements attract attention or cause frustration [9]–[11].

Together, these technologies enable a shift from static, session-based evaluations to more dynamic and real-time usability feedback. Tools like emotion-aware systems and live interaction monitors support immediate data interpretation, allowing for rapid it-

eration and timely course corrections during the design process [12]–[14]. This level of responsiveness is especially important in fast-paced development contexts, where continuous, data-driven decisions are essential [15].

Beyond speed and efficiency, AI also broadens participation in usability testing. For instance, AI-driven simulations can model the behaviors of diverse user groups or usage scenarios without requiring large participant pools. This increases both the inclusivity and generalizability of usability findings [16], [17].

However, the adoption of AI in usability testing also brings new considerations. Bastien (2010) stresses that introducing new tools must be grounded in a solid understanding of usability principles to ensure the insights remain valid and actionable [18]. Additionally, the use of AI in areas such as emotion recognition and behavioral prediction raises ethical concerns regarding transparency, privacy, and responsible data handling, as discussed by Amugongo et al. (2023) [19].

In summary, the integration of AI into usability testing represents a major advancement. By combining automation, predictive analytics, and real-time feedback, AI-supported methods address long-standing challenges related to cost, time, and scalability—while also enabling richer, faster, and more inclusive evaluations. The next section explores these techniques in more detail, focusing on the specific contributions of machine learning, NLP, and computer vision to contemporary usability testing practices.

2.3 Key AI Techniques in Usability Testing

Artificial intelligence contributes to usability testing by applying advanced algorithms to analyze extensive user data, uncover behavioral trends, and provide actionable insights that inform interface design enhancements. Among the most influential techniques are machine learning (ML), natural language processing (NLP), and computer vision. These approaches support both quantitative and qualitative evaluation, making usability testing faster, more scalable, and richer in context.

2.3.1 Machine Learning (ML)

Machine learning is widely used in usability testing to extract patterns from interaction data such as click sequences, scrolling behavior, and task completion metrics. These models help detect points of friction or user dropout that may be difficult to uncover through manual evaluation alone [6].

Supervised learning techniques can classify user actions—such as successful versus unsuccessful attempts—while unsupervised models can cluster similar behavior types, revealing less obvious usability concerns [7]. Moreover, ML can analyze historical data to anticipate usability problems early in the design process, thereby reducing the need for reactive corrections after deployment [20].

In addition to issue detection, ML supports interface personalization by adjusting features based on user preferences and behavior patterns, leading to more intuitive user experiences [21], [22]. Studies have shown that ML significantly increases the speed and scalability of usability evaluations by managing large and diverse datasets effectively [23], [24]. When integrated into AI-enabled toolkits, ML enables more comprehensive and real-time usability testing, especially in iterative, agile development cycles [25].

2.3.2 Natural Language Processing (NLP)

Natural language processing enables systems to interpret and extract insights from unstructured text sources, including open-ended feedback, survey answers, support transcripts, and user reviews. Through methods such as sentiment analysis, keyword extraction, and topic modeling, NLP helps reveal users' emotional tone, highlight recurring problems, and identify primary concerns [8], [26]–[28].

For example, sentiment analysis can indicate whether users are satisfied, frustrated, or indifferent toward specific design elements, helping teams prioritize improvements. Similarly, topic modeling can organize large sets of user comments into recurring themes, streamlining the evaluation process.

NLP also plays a role in conversational agents such as chatbots, which engage users during product interaction. These agents can gather real-time feedback, provide assistance, and reduce user confusion—making usability evaluation more continuous and em-

bedded in the user journey [29].

However, as Liang et al. (2023) caution, the usability of AI-based tools themselves can pose challenges. Complex or unintuitive interfaces and limited system feedback may hinder the effective adoption of NLP tools in practice [30]. Nonetheless, NLP and ML often work together, combining behavioral and linguistic data streams to improve the depth and accuracy of usability evaluations [25].

2.3.3 Computer Vision

Computer vision techniques analyze non-verbal cues during user interaction with digital systems. Tools such as eye-tracking are used to detect which interface components attract or lose user attention, while facial recognition software can assess emotional reactions like confusion, engagement, or satisfaction [9], [10], [31].

By examining gaze patterns, blink frequency, and micro-expressions, computer vision offers insights into users' cognitive load and focus. Gesture recognition adds an additional layer by capturing actions such as swiping or pointing, which can signal intent or difficulty navigating the interface [9], [10], [31].

This type of multimodal feedback is especially valuable in high-stakes or immersive environments, including healthcare, industrial systems, and virtual reality platforms [11], [32], [33]. Unlike ML and NLP, which process behavioral and linguistic data, computer vision focuses on interpreting visual and emotional responses, making it a critical complement in forming a complete picture of user experience.

As a whole, machine learning, natural language processing, and computer vision offer powerful methods for analyzing usability from different angles. These AI techniques not only broaden the scope of what can be measured and understood but also support the development of adaptive, real-time tools that assist teams throughout the design and testing process. The following section examines how these techniques are implemented in practical tools and frameworks designed to support usability testing in applied settings.

2.4 Tools, Frameworks, and Real-Time Capabilities in AI-Driven Usability Testing

AI technologies have significantly expanded the usability testing toolbox by introducing real-time analytics, predictive modeling, adaptive frameworks, and simulation tools. These innovations enable developers to detect usability issues during live user interactions, simulate a wider range of usage scenarios, and implement timely, targeted design improvements. This section discusses major tool categories, including real-time analytics platforms, predictive systems, and AI-enhanced user interface frameworks.

2.4.1 Real-Time Data Collection and Analysis

AI-powered tools can collect and process user interaction data in real time, providing immediate insights to designers and developers. This functionality is particularly beneficial in agile development environments, where rapid adaptation to user feedback is essential [13], [14]. Real-time feedback loops help identify problems early and reduce development bottlenecks, leading to faster design iterations.

These tools can highlight areas of potential failure and direct developer attention to critical interface components, improving both efficiency and user experience outcomes [15]. For example, Drungilas et al. (2024) [12] integrated emotion tracking into usability tools, offering insights into users' emotional states during interaction. This affective data deepens the understanding of user experience and helps refine design decisions more precisely.

Additionally, AI enhances risk detection within development workflows, enabling agile teams to act proactively [34]–[36]. Advanced data tools also contribute to managing innovation pipelines by improving responsiveness in dynamic industrial settings [37].

However, as Bastien (2010) emphasizes, understanding the foundations of usability methods remains crucial—even as AI enables faster, scalable feedback [18]. Ethical considerations are equally important; Amugongo et al. (2023) underscore the need for transparent and responsible AI use to maintain fairness and trust in evaluation practices [19].

2.4.2 Scalability and Predictive Analytics

Artificial intelligence also contributes significantly to scalability by automating usability testing tasks across large and diverse user populations. Large user interface models (LUIMs) and behavior simulation techniques allow developers to assess system performance without extensive participant recruitment [16]. AI systems can simulate varying user behaviors, reducing logistical constraints and supporting more flexible, continuous testing [17].

Predictive analytics tools further extend these capabilities by forecasting potential usability issues based on current and historical interaction data. By identifying likely failure points before deployment, teams can proactively adjust designs and reduce the cost of post-release corrections [37]–[39]. This predictive functionality enhances both the efficiency and quality of digital products.

As part of this trend, AI-based evaluation tools like the Chatbot Usability Scale demonstrate how large-scale systems can maintain a user-centered focus. These tools assess the interaction quality of AI-powered conversational agents, emphasizing usability and communication effectiveness [27].

In addition, Kang and Lou (2022) [40] discuss the interplay between human and AI decision-making in usability testing. Their findings highlight how shared judgment between systems and evaluators enhances engagement and enables more effective optimization across platforms.

2.4.3 AIMT-UXT and AI-driven Tools

Artificial intelligence and mouse tracking-based user experience tool (AIMT-UXT) [41] is a notable example of an AI-enhanced usability testing tool that combines real-time analytics with adaptive feedback. The tool integrates mouse tracking data to analyze user interactions on websites—particularly in the public sector—recording movements, clicks, and hovers to detect interface confusion or inefficiency.

A fuzzy inference system within AIMT-UXT is used to handle uncertainty in behavior data, helping classify usability performance at the element level. This functionality enables the identification of problematic areas that require design improvement.

The tool also applies unsupervised learning techniques to cluster user behavior patterns. This grouping helps reveal widespread usability issues affecting varied user types. A major advantage of AIMT-UXT is its immediate feedback capability, allowing issues to be addressed during live sessions, thereby accelerating design iterations.

Additionally, the system features adaptive testing, which dynamically adjusts in response to user behavior—ensuring that core challenges are addressed efficiently.

According to Boodaghian Asl et al. (2024), combining AIMT-UXT with hybrid modeling approaches can improve its simulation and classification capabilities. These enhancements may result in more accurate behavior predictions and better-informed interface revisions [41]–[43].

2.5 Case Studies and Applications

Several case studies highlight the practical impact of AI in usability testing. These studies demonstrate AI's potential to provide more detailed and actionable insights, ultimately improving the overall user experience.

2.5.1 Intelligent Medical Systems

Artificial intelligence (AI) has practical applications in intelligent medical systems to improve efficiency, accuracy, and user interaction in healthcare. The study by Donghong Zhou [9] on AI-driven interfaces demonstrates that AI can streamline data processing and enhance human-computer interaction (HCI) within these systems. According to the research, task completion times were reduced by an average of 24.6%, with specific tasks, such as downloading job information, achieving up to 65.84% improvement in efficiency. These results highlight AI's role in reducing the cognitive load on healthcare professionals, allowing for more intuitive operation and quicker decision-making under pressure. The paper also underscores the importance of designing interfaces that integrate effectively with the workflows of medical professionals, making their tasks more manageable. Furthermore, the AI system in this study facilitates real-time data collection and processing of physiological parameters, including blood pressure, ECG, and pulse wave signals, which are transmitted to a smartphone application via a WiFi module. This integration allows

healthcare professionals to access critical patient data efficiently, supporting timely and informed decision-making as the system captures and analyzes essential metrics in real time.

2.5.2 AI-Driven Design Heuristics in UX Development

The paper by Jin et al.(2021) investigates how AI can improve UX design by developing design heuristics that assist designers in creating AI-driven concepts. Through an analysis of over 1,700 AI patents, the authors extracted 40 design heuristics aimed at inspiring UX designers during the conceptual design phase. These heuristics are shown to expand the design space, enabling the creation of innovative and user-centered AI applications. The study emphasizes that understanding AI capabilities and incorporating them into the early stages of design can lead to more effective and creative solutions, highlighting the potential of AI in defining the future of UX [44].

2.5.3 Smart Online Shopping Systems

In smart online shopping systems AI is used for evaluating product quality, providing personalized recommendations, and enhancing security features. Machine learning algorithms examine user behavior and preferences to provide a personalized and secure shopping experience which is known to improve user satisfaction and increase the likelihood of repeat purchases [45]. For instance, deep neural collaborative filtering (DNCF) models can effectively produce personalized recommendations by analyzing user interaction data, yielding a precision rate of 85% and a recall rate of 78%—indicators of the model's strong capability to deliver relevant suggestions in e-commerce environments [46]. Additionally, research on The study on Explainable AI (XAI) by [47] illustrates how scenario-based approaches can be used to uncover user-specific explanation needs, especially within the context of fraud detection. This study developed scenarios such as "Clear Transaction Fraud" and "Uncertain Transaction Fraud," which reveal the critical explanation elements needed to reduce cognitive load and prevent user overload. These scenario-based requirements enhance both trust and usability by ensuring that explanations are selectively informative and aligned with user needs, developing a more user-centric and secure shopping environ-

ment.

2.5.4 AI in Various Industries

AI integration across various industries has become trans-formative, including in enhancing usability testing and user experience. AI-powered usability testing tools have been employed across diverse sectors, ranging from healthcare to finance, automating routine tasks and providing insights into user interactions.

Healthcare Sector The application of artificial intelligence (AI) in healthcare continues to support improvements in both operational efficiency and clinical decision-making. Studies have demonstrated how AI-driven systems can streamline workflows and reduce cognitive load for healthcare professionals, enabling quicker and more accurate responses in clinical environments [9]. Similarly, Karalis (2024) reviewed AI's broader role in clinical practice, noting its ability to analyze extensive datasets to support diagnostic accuracy and personalized treatment plans across specialties such as cardiology and gastroenterology. This review underscores the potential of AI to improve patient outcomes and alleviate clinician workload, while also stressing the need to address ethical considerations—such as data privacy and transparency—as AI becomes more deeply embedded in healthcare environments [48].

Finance Sector In the finance industry, AI-driven usability testing tools are employed to streamline user interfaces for banking apps and financial services platforms. By analyzing user interactions and behavior, these tools contribute to the design of more intuitive and user-centered interfaces, ultimately enhancing user satisfaction and fostering greater trust in digital financial services [49].

Retail and E-commerce AI is extensively utilized in retail and e-commerce to optimize customer experiences through sophisticated personalization techniques. Specifically, deep neural collaborative filtering approaches are employed to analyze user data, including previous purchases and browsing patterns, enabling the delivery of highly personalized product recommendations. These AI-driven approaches improve the relevance of product recommendations and play a key role in optimizing the overall usability of e-commerce platforms by aligning them closely with individual customer preferences. This tailored

approach leads to improved customer satisfaction and increased engagement [46].

These case studies and industry applications underscore AI's significant potential for improving usability across various domains. By providing more detailed and actionable insights, AI enhances the efficiency and effectiveness of usability evaluations and consequently can improve the overall user experience. As AI technologies progress, their incorporation into usability testing is set to become increasingly influential, opening up new opportunities for innovation and enhancement in user experience design.

2.6 Recent Advancements in AI for Usability Testing

Recent advancements in AI for usability testing have been driven by the rapid development of generative AI, multi-modal models, and intelligent automation. These technologies have transformed the way usability testing is conducted.

2.6.1 Generative AI and Multi-modal Models

Generative AI, exemplified by models developed by OpenAI and Google DeepMind, has advanced the field of usability testing. These models are capable of generating realistic and diverse user scenarios, which can be used to simulate a wide range of interactions with digital interfaces. By creating synthetic data that mimics real user behavior, generative AI can help in identifying potential usability issues before they affect actual users [50].

Multi-modal models can capture and analyze complex user interactions. These models combine data from visual, textual, and behavioral sources to assess how users interact with systems that require diverse inputs, such as voice commands, gestures, and visual feedback. According to Dibeklioglu et al.(2021) [51], multi-modal analysis methods integrate tools across different modalities, allowing for a more comprehensive evaluation of user behavior in varied contexts. For instance, the editorial discusses studies that use multi-modal datasets to assess user experience during interactive scenarios, such as personality analysis from audiovisual cues and pain assessment from body movements using long short-term memory (LSTM) networks. In each case, these models provide a holistic view by processing information from multiple channels simultaneously, which is especially valuable in complex systems where user inputs are dynamic and varied. By combining

multiple data sources, this method offers a deeper insight into user experience, revealing nuanced interaction patterns that might go unnoticed with isolated data analysis. As a result, it helps creating interfaces that better match the ways people actually use them in real-world situations.

2.6.2 Intelligent Automation and Digital Twins

Virtual representations known as digital twins are increasingly utilized to simulate and study user interactions in controlled settings, enabling detailed analysis of system behavior. According to Cichon and Rossmann (2017), simulation-driven interfaces within digital twins support testing before, during, and after operations, allowing real-time scenario evaluation and usability analysis without requiring physical interaction. This method is particularly beneficial for evaluating complex systems like robotics, where virtual interactions can identify potential usability issues early, improving safety and design [52]. In their systematic review, Barricelli and Fogli (2024) explore the role of digital twins in human-computer interaction (HCI), highlighting that despite their strong real-time functionalities, their effectiveness is often constrained by insufficient emphasis on user-centered design. The review highlights a gap in accessibility for non-expert users, recommending that future digital twin designs prioritize intuitive interfaces and user engagement across varied industries [53].

In fault diagnosis, digital twins extend beyond usability testing by simulating operational states that may lack sufficient real-world data. Xia et al.(2021) proposed a digital twin framework for fault diagnosis, integrating deep transfer learning to improve fault detection even with limited data. This model achieved high diagnostic accuracy by continuously updating with real-time operational data, making it particularly effective in applications like triplex pump fault detection [54]. Xu et al.(2019) further demonstrated that digital twins, when paired with high-fidelity data simulations, enhance fault detection in smart manufacturing systems, addressing challenges related to data scarcity [55]. Segovia and Garcia-Alfaro (2022) emphasize digital twins' role in predictive maintenance through continuous monitoring, allowing for proactive interventions that support system health and extend operational lifespan [56]. Meanwhile, Zayed et al.(2023) achieved up

to 96.8% accuracy in fault detection within critical industrial applications by combining digital twins with optimized machine learning models, underscoring their effectiveness in improving system reliability and performance [57].

Real-time analytics further augment these capabilities by enabling continuous monitoring and immediate detection of faults. For example, Hodavand et al.(2023) demonstrate how digital twins in HVAC (Heating, Ventilation, and Air Conditioning) systems use real-time analytics to identify and mitigate faults early, reducing energy consumption and enhancing operational efficiency [58]. Wang et al.(2023) extend this approach with a digital twin model for bearing fault diagnosis, combining numerical simulation and machine learning to provide real-time insights that guide maintenance decisions, thus increasing reliability and preventing potential system failures [59]. These applications show the potential of digital twins, particularly when enhanced with real-time analytics, to improve diagnostic precision, operational efficiency, and proactive maintenance across diverse sectors. Despite their promise, digital twin systems and other advanced AI tools can be challenging to develop and maintain. As Islam et al. (2023) point out, training and fine-tuning AI models demand substantial technical expertise and ongoing resources to ensure their continued accuracy and relevance [60].

2.7 Model Transparency and Ethical Considerations in AI-Driven Usability Testing

With the growing integration of artificial intelligence into usability testing, ensuring model transparency has become essential for promoting effectiveness, building user trust, and supporting the ethical use of these technologies. Understanding how AI tools generate insights is essential for designers to make informed decisions and ensure fairness in evaluation practices. Moreover, when AI-generated usability improvements are implemented, end users are directly affected. Transparency in how these insights are produced helps build user trust, acceptance, and confidence in the final design.

Transparency for Informed Design Liao et al. (2023) [61] investigated the real-world needs of UX designers engaging with pre-trained AI models and highlighted that

transparency plays a key role in helping teams grasp system functionality, such as the nature of training data, model capabilities, and the underlying logic behind decisions. This level of clarity allows designers to interpret AI-generated outputs with greater confidence and to adapt interface designs accordingly.

In contexts where AI systems perform classification or prediction tasks, knowing where and why a model may fail helps designers set appropriate constraints and expectations. Rai (2020) and Doshi-Velez & Kim (2017) similarly argue that explainable AI (XAI) enables users to better grasp model behavior and avoid blindly accepting outputs that could be biased or erroneous [62], [63].

Building User Trust and Acceptance Transparent AI models also enhance user trust—especially in usability testing environments where the system's influence may affect interface design decisions. When UI designer and stakeholders understand how and why AI systems reach certain conclusions, they are more likely to accept recommendations and feel secure in their use. Amershi et al. (2019) support this view by providing guidelines for human-AI interaction design that emphasize interpretability and clear communication of AI capabilities and limitations [64].

Fan et al. (2024) and Liao et al. (2024) [65][66] further highlight that transparent systems are more likely to support collaborative decision-making and iterative refinement, as stakeholders are empowered to question, confirm, or reject AI findings based on evidence.

Ethical AI Practices As AI tools take on a greater role in user research, ethical issues—including bias, accountability, and data privacy—are becoming increasingly significant and demand careful consideration. Transparent systems support ethical practices by making it easier to detect and correct unintended consequences. For example, access to model training data and processing logic can reveal whether certain user groups are underrepresented or unfairly treated in usability outcomes.

The 2024 AI Index Report from Stanford University identifies explainability as a key requirement for responsible AI implementation across domains, including education, healthcare, and user experience research [25]. Haque, Islam, and Mikalef (2023) [67] also stress that transparency is essential for bridging the interpretability gap between complex AI systems and non-expert users—helping ensure responsible deployment and informed

consent in data-driven environments.

Nevertheless, the increasing complexity of AI systems brings the risk of reinforcing biases embedded in their training data, which can result in distorted usability findings or design choices that marginalize certain user groups [68]. Ensuring transparency is therefore essential not only for ethical accountability but also for identifying and correcting such biases early in the design process.

Supporting Collaboration and Interdisciplinary Teams AI transparency also enhances collaboration between UX researchers, designers, and technical developers. Yu et al. (2024) explain that shared understanding of an AI model's structure and logic helps cross-functional teams define common goals, align priorities, and jointly interpret results [69]. This collaboration is vital in AI-supported usability testing, where design outcomes often depend on both algorithmic feedback and human judgment.

In high-stakes environments, where usability decisions can significantly impact safety, accessibility, or public trust, the ability to interpret AI outputs transparently is not just a convenience—it is a necessity. Without transparency, AI tools risk becoming "black boxes" that hinder accountability and compromise user experience quality.

In summary, transparency is central to the ethical and effective integration of AI in usability testing. It fosters trust, supports informed design decisions, enables collaboration, and ensures that AI-generated insights remain interpretable and actionable. As usability testing evolves to include more AI-driven methods, maintaining model transparency will be essential for preserving both methodological rigor and user-centered values.

2.8 Comparison of AI and Human Usability Testing

AI-driven usability testing offers distinct advantages over traditional human-led methods, particularly in terms of speed, scale, and consistency. However, human evaluators continue to play a crucial role in interpreting user behavior, emotions, and contextual nuances—elements that AI alone may misinterpret or overlook. This section highlights the comparative strengths of each approach and supports the case for integrated, hybrid methods.

2.8.1 Complementary Strengths

Artificial intelligence excels at processing large volumes of data quickly and accurately. It can analyze user behavior logs, detect patterns, and provide immediate feedback during development—capabilities that significantly accelerate usability cycles. For instance, AI systems can evaluate user interaction data across diverse user profiles without the need for extensive participant recruitment, making usability testing more scalable and inclusive [70].

Tools equipped with real-time analysis features allow developers to detect problems during active sessions and adjust designs rapidly. This contrasts with human-led testing, which typically requires post-session analysis and manual interpretation, often delaying actionable feedback [27], [71]. Moreover, as Sharma et al. (2022) notes, AI systems can use behavioral modeling to flag potential interface issues before they escalate or reach users, thus supporting early-stage design refinement [72].

However, AI-based analysis is limited in its ability to interpret emotional nuance, subjective satisfaction, and complex intent. Human evaluators, on the other hand, can contextualize behavior, understand user narratives, and infer emotional responses—essential components of holistic usability testing. For example, asynchronous remote usability testing studies have shown that human moderation allows users to feel more at ease, often leading to the discovery of content and navigation issues that automated systems might miss [73].

2.8.2 Human-AI Integration in Practice

Recent research supports a collaborative model where AI and human evaluators work together. Gannouni et al. (2023) demonstrated that while AI systems using EEG data could detect emotional states, human input was needed to interpret the context behind those emotions and connect them to usability satisfaction [74].

Similarly, Kuang et al. (2023) discovered that user preferences varied when engaging with AI-driven conversational assistants, with text interfaces favored for rapid responses and voice interfaces preferred for more natural, conversational interactions. This user preference highlights the role of human oversight in deciding how and when to deploy

specific AI modalities [75].

In scenario-based testing, van Eck et al. (2018) showed that while AI could generate realistic test scenarios from behavioral data, it lacked the ability to explain anomalies or prioritize findings. Human reviewers were essential for interpreting the significance of outlier behavior and providing contextual recommendations [76].

The integration of AI in software development tools also reflects this dynamic. For instance, GitHub Copilot uses AI to generate code suggestions, but developers often need to adjust or reject outputs to suit specific project contexts—underscoring the importance of human judgment even in technically precise environments [30].

Research by Bansal et al. (2021) warns that while explanations from AI can build user trust, they can also lead to over-reliance. Human evaluators act as a safeguard by critically assessing AI recommendations and ensuring that users are making informed decisions [77]. Likewise, Radziwill and Benton (2017) argue that chatbot evaluations must be supported by human reviewers, particularly in scenarios involving unexpected queries or diverse user needs [78].

In conclusion, the comparison between human and AI usability testing is not a matter of replacement but of augmentation. AI systems bring speed, scale, and consistency, while human evaluators contribute depth, empathy, and contextual understanding. Together, these complementary strengths form the basis of a hybrid usability testing model—one that leverages the efficiency of AI without sacrificing the interpretive richness of human insight.

2.9 Synthesis and Identified Gaps

The reviewed literature illustrates the growing role of artificial intelligence in transforming usability testing practices. AI technologies—particularly machine learning, natural language processing, and computer vision—enable faster, more scalable, and data-rich evaluation processes. These tools offer real-time feedback, simulate user behavior, detect patterns, and even predict potential usability issues before deployment. AI-supported tools such as AIMT-UXT, digital twins, and emotion-aware systems further extend testing capabilities by integrating automation, adaptability, and predictive analytics.

However, despite these advancements, AI-based usability testing is not without limitations. While AI systems are proficient at pattern detection and data analysis, they often lack the contextual understanding, empathy, and interpretive depth that human evaluators provide. Human expertise remains essential for identifying emotionally driven usability problems, understanding user intent, and making nuanced design judgments. Moreover, ethical concerns such as model transparency, algorithmic bias, and over-reliance on automated systems highlight the importance of responsible and explainable AI design.

The literature strongly supports a hybrid approach—one that combines AI's speed, consistency, and scale with human insight, interpretive reasoning, and ethical oversight. Such a model not only leverages the operational strengths of AI but also ensures that usability evaluations remain human-centered and adaptable to varied user contexts.

Despite these promising developments, notable gaps remain. First, existing studies tend to focus on specific industries, tools, or techniques in isolation, without a clear framework for comparing AI-based and human-led evaluations in a systematic, task-specific context. Second, many studies lack empirical comparison of AI-generated evaluations with human assessments, especially within heuristic usability testing scenarios. Finally, limited research has been conducted on the practical integration of large language models (LLMs) like GPT-4 into real-time, web-based usability analysis tools—particularly in accessible domains such as public utility websites.

This thesis addresses these gaps by developing and implementing a custom GPT-4o-integrated API designed for heuristic usability evaluation of online unit converter websites. It systematically compares AI-generated results with those from human evaluators to assess alignment, divergence, and areas of complementarity. By doing so, the research aims to propose a practical, hybrid framework that balances AI's efficiency with human judgment—advancing usability testing toward more responsive, reliable, and user-centered design processes.

3 Materials and methods

3.1 Experimental design

This study analyzes the capabilities of AI-driven usability testing by comparing its performance with that of human evaluators in identifying usability issues. The methodology includes material selection, task design, data collection, and comparative analysis.

3.1.1 Participant selection and task design

Two evaluation groups, AI-based analysis and human student evaluations, were used. The human evaluators were students from the "Usability, User Experience, and Analytics 2022 - DTEK0069" course at the University of Turku. Each student selected and analyzed three online unit converters from a given list. They were asked to apply any five of Nielsen's ten usability heuristics that seemed appropriate to them. The task given to the users involved converting between feet and meters to assess tool usability. The following unit converters were suggested for the evaluation:

- **UnitConverters.net**: `https://www.unitconverters.net/length-converter.html`
- **Convert-me.com**: `https://www.convert-me.com/en/convert/length/`
- **Omni Calculator**: `https://www.omnicalculator.com/conversion/length-converter`
- **Asknumbers**: `https://www.asknumbers.com/`
- **The Engineering Toolbox**:
  `https://www.engineeringtoolbox.com/length-units-converter-d_1033.html`

3.1.2 Exclusion of omni calculator

Omni Calculator was excluded due to a significant interface redesign that occurred before the study. Since student data was collected after the course had finished and the study was conducted at a later time, a considerable gap existed between these phases. During this period, the website underwent a major interface update. As both AI and human evaluations required a consistent interface for valid comparison, Omni Calculator was removed from the analysis to maintain data integrity.

3.1.3 Data collection and privacy

Thirty student reports were analyzed. They were anonymized, ensuring no personal information was included. Students assigned severity ratings from 0 to 4 and proposed solutions for identified usability issues. Their evaluations were compiled into comparison reports, ranking unit conversion tools based on usability performance. The AI evaluation was conducted using screenshots uploaded to an API built with OpenAI's GPT-4o model. While GPT-4o is not explicitly designed for usability testing, it was used in this study to assess interface elements against Nielsen's heuristics. The model generated heuristic evaluations based on textual and visual patterns within the screenshots. The severity of each issue was on a scale from 0 to 4, where:

- **0** — No usability problem
- **1** — Cosmetic problem
- **2** — Minor usability problem
- **3** — Major usability problem; important to fix
- **4** — Usability catastrophe; imperative to fix

3.1.4 Comparison and analysis process

Usability feedback from both AI and student evaluations was systematically collected and analyzed. The analysis focused on identifying recurring usability issues, categorizing them

according to heuristic principles, and comparing the average severity ratings assigned by AI and human evaluators. Where applicable, statistical analysis was applied to quantify trends and highlight key differences. The comparison helped reveal AI's effectiveness in detecting usability issues and how its evaluation differed from human judgment.

3.2 API design and implementation

This study utilized an API that was custom-built to conduct usability evaluations based on image analysis, optimizing OpenAI's GPT-4o model. GPT-4o was selected due to its advanced natural language processing (NLP) capabilities, ability to analyze extracted text effectively, and enhanced contextual understanding. Compared to previous versions, GPT-4o provides improved response consistency and better handling of heuristic-based usability evaluations. The API, developed in Flask, allows users to upload images, which are processed to extract textual content before being analyzed for usability issues. The following section details the API's core functionalities, including image processing, interaction with OpenAI's API, error handling, and response management.

3.2.1 Image upload and conversion

The API features a web-based interface that supports image uploads in JPEG and PNG formats. Upon upload, the API processes images using Optical Character Recognition (OCR) via the ***pytesseract*** library to extract relevant textual content for analysis.

This code snippet in Listing 3.2.1 shows how this API handles an image upload, instantly converts it into text, and prepares the extracted text for further analysis.

3.2.2 API call to OpenAI

Once the text is extracted, the API constructs a request payload and sends it to OpenAI's GPT-4o model for heuristic-based usability evaluation. The model assesses the extracted content against Nielsen's all ten usability heuristics, identifying usability issues and providing structured feedback.

This snippet in Listing 3.2.2 shows how the API constructs the request payload and

```
@app.route('/analyze-image', methods=['POST'])
def analyze_image():
    if 'files[]' not in request.files:
        return jsonify({'error': 'No file part'}), 400

    files = request.files.getlist('files[]')
    if not files:
        return jsonify({'error': 'No selected file'}), 400

    if len(files) > 16:
        return jsonify({'error': 'You can upload a maximum of 16 images.
            '}), 400

    try:
        extracted_texts = []
        for file in files:
            image = Image.open(file)
            extracted_text = pytesseract.image_to_string(image)
            extracted_texts.append(extracted_text)

        combined_text = "\n\n".join(extracted_texts)
```

Listing 3.2.1: Handling image upload and conversion.

interacts with the OpenAI API to fetch the usability analysis. The API evaluates usability based on static screenshots, which limits its ability to assess real-time interactions and dynamic feedback. This aspect is further discussed in the Discussion chapter.

3.2.3 Error handling and response management

To ensure reliability, the API implements error-handling mechanisms to manage issues such as invalid API responses, image processing failures, and exceeded file upload limits. If an error occurs, the API logs errors with detailed messages and returns user-friendly responses, ensuring transparent troubleshooting and maintaining system robustness, as shown in Listing 3.2.3.

3.2.4 Analysis and response delivery

If multiple images are uploaded, the API extracts text from each and combines the results into a unified analysis report, ensuring consistency in usability evaluation. The report is formatted in markdown for readability. The response is structured to highlight key usability concerns, categorize heuristic violations, and provide actionable recommendations [Listing 3.2.4].

```
headers = {
    "Content-Type": "application/json",
    "Authorization": f"Bearer {openai_api_key}"
}

combined_analysis_payload = {
    "model": "gpt-4o",
    "messages": [
        {
            "role": "user",
            "content": f"""
            Review the following conversion online tool:
Write a report about the converter based on a heuristic evaluation using
    the 10 Nielsen heuristics.....................
            **Extracted Text**:
            {combined_text}
            """
        }
    ],
    "max_tokens": 2000
}

combined_analysis_response = requests.post("https://api.openai.com/v1/
   chat/completions", headers=headers, json=combined_analysis_payload)
```

Listing 3.2.2: Making an API call to OpenAI.

```
if combined_analysis_response.status_code != 200:
    return jsonify({'error': 'Failed to get valid response from OpenAI',
        'details': combined_analysis_response.text}),
       combined_analysis_response.status_code
```

Listing 3.2.3: Error handling during an API call.

3.2.5 Follow-up questions

Users can submit follow-up queries related to the initial analysis. The API maintains conversational context by linking follow-up responses to previous analyses, enabling iterative usability exploration. This functionality enables deeper exploration of usability concerns and possible improvements based on iterative feedback [Listing 3.2.5].

3.3 Data analysis

The following section presents an analysis of usability issues detected by both AI and human evaluators across four websites: UnitConverters.net, Engineering Toolbox, Convert-Me, and AskNumbers. The results are examined based on heuristic principles, with a focus on identifying patterns, similarities, and differences between the two sources of eval-

```
combined_analysis = combined_analysis_response.json()['choices'][0]['
    message']['content']
combined_analysis_html = markdown(combined_analysis, extras=["tables","
    fenced-code-blocks", "strike", "break-on-newline"])

conversation_id = str(uuid.uuid4())
conversations[conversation_id] = [combined_text, combined_analysis]

return jsonify({'conversationId': conversation_id, 'combinedAnalysis':
    combined_analysis_html})
```

Listing 3.2.4: Combining and formatting the analysis.

uation. The analysis utilizes visual representations such as bar charts, heatmaps, and pie charts to support the findings.

3.3.1 Clarification on Severity Ratings

In the following tables, average severity ratings are presented for each heuristic per website, based on AI and student evaluations. However, these averages do not indicate the number of issues found in each heuristic category. Some categories may include only one high-severity issue, while others may have several lower-severity issues. Therefore, the average values should be interpreted as indicative rather than definitive representations of overall usability severity.

3.3.2 Website 1 analysis: Unit Converter

The evaluation of UnitConverters.net revealed critical usability concerns that impacted both system functionality and user experience. While AI assessments primarily focused on structural and technical inconsistencies, student evaluations highlighted challenges related to the workflow efficiency and ease of use. Despite their differing perspectives, both evaluators identified overlapping problems that hindered usability.

One of the most pressing concerns was the lack of immediate feedback when users entered values. AI flagged this issue as a high-severity concern (3), noting that the absence of loading indicators or confirmation messages made the system appear unresponsive. Students recognized this frustration, emphasizing that conversion results often blended into the interface, making them difficult to notice. To address this, AI recommended integrating real-time validation and loading animations, while students suggested repositioning

```
@app.route('/ask-follow-up', methods=['POST'])
def ask_follow_up():
    data = request.json
    question = data.get('question')
    conversation_id = data.get('conversationId')

    if not question or not conversation_id:
        return jsonify({'error': 'Missing question or conversationId'}),
            400

    if conversation_id not in conversations:
        return jsonify({'error': 'Invalid conversation ID'}), 400

    previous_messages = conversations[conversation_id]

    try:
        headers = {
            "Content-Type": "application/json",
            "Authorization": f"Bearer {openai_api_key}"
        }

        messages = [{'role': 'user', 'content': message} for message in
           previous_messages]
        messages.append({'role': 'user', 'content': f"Please answer the
           following question based on the previous analyses: {question}
           "})

        payload = {
            "model": "gpt-4o",
            "messages": messages,
            "max_tokens": 2000
        }
        follow_up_response = requests.post("https://api.openai.com/v1/
           chat/completions", headers=headers, json=payload)
```

Listing 3.2.5: Handling follow-up questions.

the results for better visibility.

Another significant usability barrier was the absence of an undo/reset function, restricting user control and flexibility. AI rated this as a severe issue (4), as users were forced to manually clear inputs, increasing interaction effort. Students also identified this limitation but focused more on the inconvenience of re-entering values, assigning it a slightly lower severity score (3). While AI proposed a dedicated reset button, students emphasized the need for a "reverse units" button, which would allow seamless swapping between input and output fields.

The cluttered interface and excessive advertisements further disrupted usability. AI categorized this as a usability catastrophe (4), citing poor spacing, redundant elements, and an overwhelming layout that distracted users. Students shared similar concerns

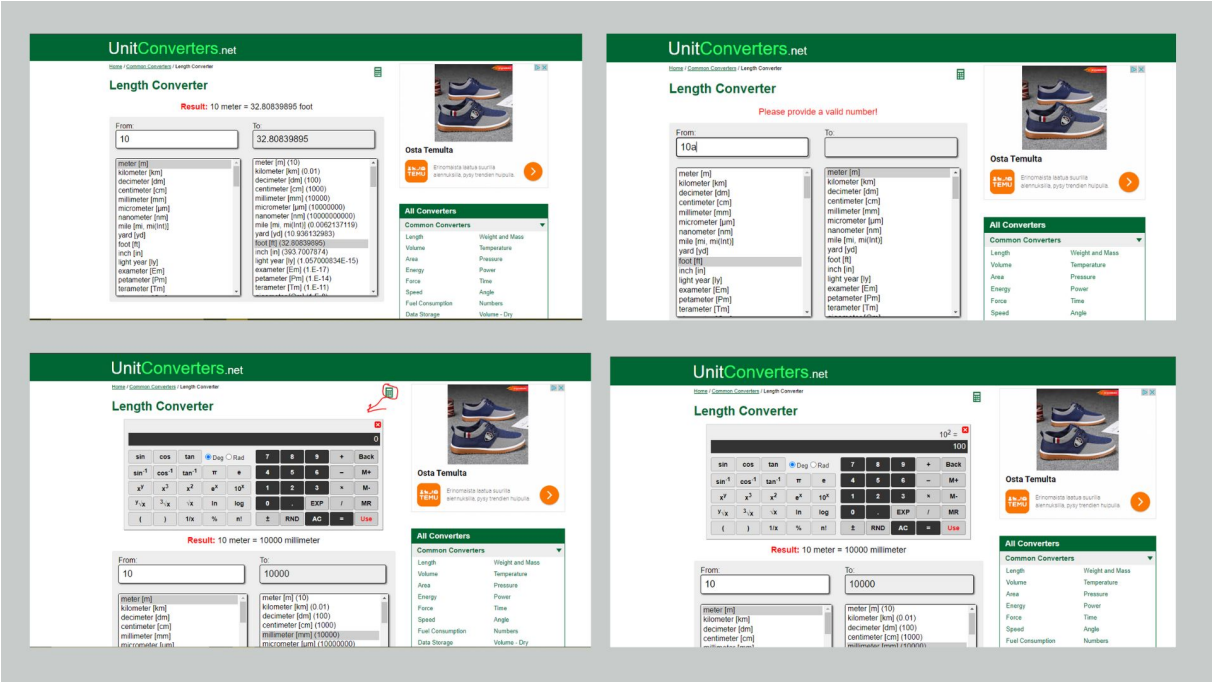

Figure 3.1: Interface of the UnitConverters.net website (screenshot by the author).

but focused more on the placement of advertisements, which they found intrusive and workflow-disrupting. While AI recommended simplifying the design to enhance readability, students proposed restructuring the layout by grouping related conversion categories into collapsible sections.

Another major issues were weak input validation and unclear error handling. AI identified that the system permitted non-numeric inputs, leading to failed conversions without proper warnings (3). Additionally, AI pointed out that vague error messages, such as "Please provide a valid number!", did not offer actionable guidance. Students similarly struggled with this issue, specifically noting that the converter failed to recognize regional decimal formats (dots vs. commas), resulting in input errors. Both evaluators suggested improvements: AI recommended stricter validation mechanisms and clearer error prompts, while students proposed allowing flexible input formats to accommodate different user conventions.

Navigation inefficiencies further compounded usability challenges. AI noted that the absence of a search function forced users to scroll through long drop-down lists to locate desired units (2). Students viewed this as an even greater problem (3), highlighting the tedious nature of manually searching for units. AI recommended refining unit categorization to streamline selection, whereas students proposed implementing a search bar to

improve accessibility.

Inconsistencies in formatting and terminology were also apparent. AI flagged unit labels as inconsistent in style and abbreviation usage (2), which could cause confusion. Students similarly pointed out discrepancies but were more concerned about how non-standard terminology affected real-world comprehension. To address this, AI suggested standardizing unit names and abbreviations, while students recommended providing inline explanations for lesser-known units.

Another usability challenge stemmed from poorly designed error messages. AI rated this as a severe issue (4) due to generic and unhelpful feedback that failed to guide users toward a solution. Students encountered similar difficulties but emphasized the lack of instructional content within the messages, making it unclear how to correct input errors. AI proposed making messages more detailed and informative, while students suggested placing them closer to the input fields to enhance visibility.

Limited user control options further hindered efficiency. AI identified the inability to modify past inputs or save frequently used conversions as a usability flaw (3). Students focused on a related concern, noting that there was no quick swap function to reverse unit selections. AI recommended implementing a conversion history feature, while students emphasized the need for a dedicated swap button.

Lastly, help and documentation deficiencies were evident. AI gave this issue a low severity rating (1), viewing it as a minor shortcoming due to the absence of tool-tips or in-page explanations. Students, however, found this to be a more significant problem (3), as there were no clear descriptions of units or an FAQ section to assist with unfamiliar measurements. AI recommended adding hover tool-tips, while students suggested incorporating a searchable help section for better accessibility.

The usability issues in UnitConverters.net highlight a blend of technical limitations and interaction inefficiencies. AI evaluations predominantly focused on structural refinements, such as error prevention, layout consistency, and system feedback, whereas students emphasized workflow-related challenges, such as navigation barriers and visibility problems. Addressing these concerns through real-time input validation, improved feedback mechanisms, streamlined navigation, and clearer error messaging would greatly

enhance the overall user experience.

To better illustrate these findings, Table 3.1 presents a comparative summary of the severity ratings assigned by both evaluators, while Figure 3.2 provides a visual representation of the discrepancies between AI and student assessments.

Table 3.1: Comparison of average severity ratings between AI and students evaluations for the *UnitConverters.net* website.

Heuristic	**AI Severity**	**Student Severity**
Visibility of System Status	2.75	1.50
Match Between System and Real World	2.00	2.50
User Control and Freedom	3.00	2.29
Consistency and Standards	2.00	2.20
Error Prevention	2.50	2.31
Recognition Rather Than Recall	2.00	1.67
Flexibility and Efficiency of Use	1.00	2.15
Aesthetic and Minimalist Design	2.50	1.67
Help Users Recognize, Diagnose, and Recover	4.00	2.75
Help and Documentation	1.50	2.67

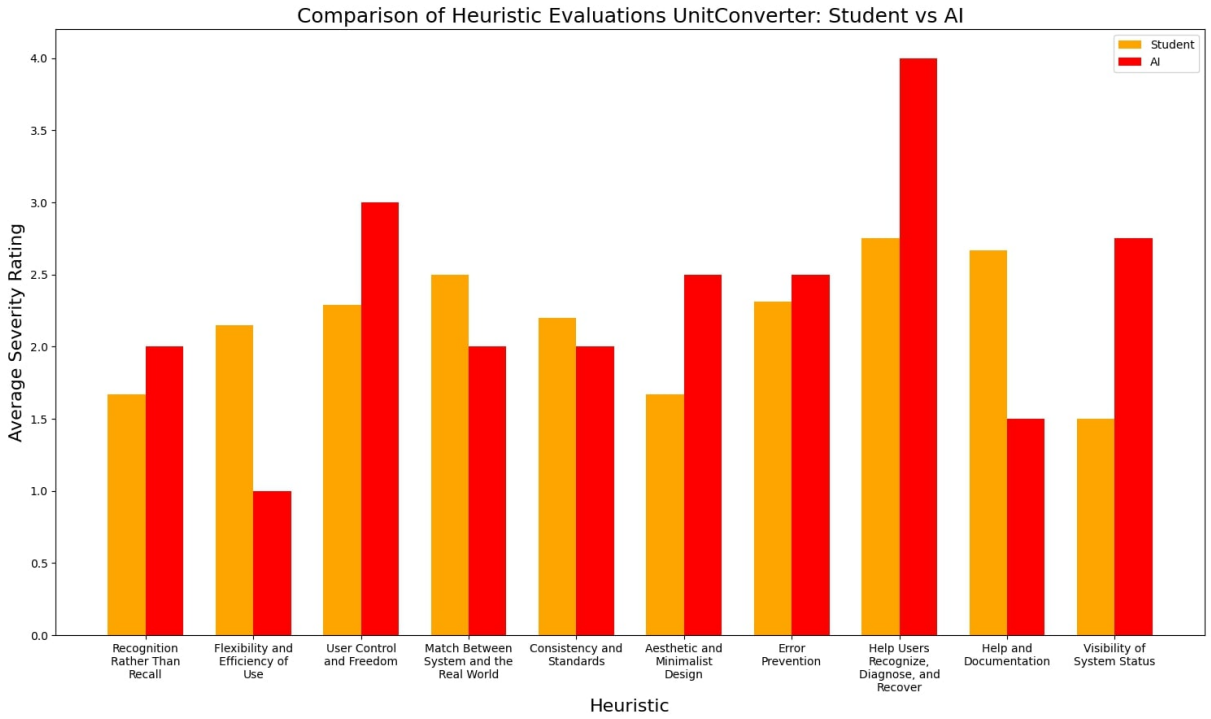

Figure 3.2: Comparison of Heuristic Evaluations for UnitConverters.net: AI vs. Students.

3.3.3 Website 2 analysis: Engineering toolbox

While the first website suffered from visibility issues and workflow inefficiencies, Engineering Toolbox presented obstacles related to navigation complexity, unclear input validation,

and the absence of system feedback. AI's assessment focused on structural inconsistencies and missing validation mechanisms, while students emphasized the practical difficulties users faced in interacting with the tool.

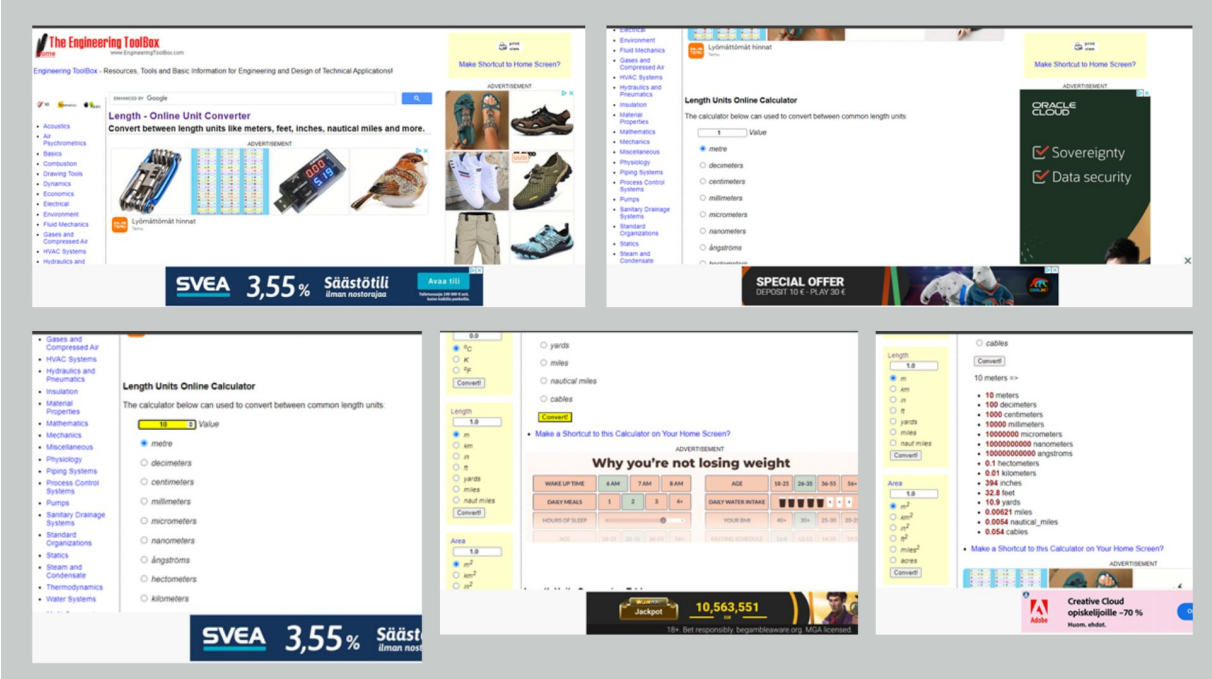

Figure 3.3: Interface of the EngineeringToolbox website (screenshot by the author).

A major concern identified in this evaluation was the difficulty in locating conversion results. AI noted that the system lacked clear feedback indicators, making it difficult for users to determine whether their input had been processed. Students also had trouble with the conversion result, but they explained it in a different way—they said the results weren't placed where they expected, which made their work harder. While AI recommended adding real-time system responses, such as confirmation messages and loading indicators, students suggested repositioning the conversion results to improve visibility.

Another usability challenge stemmed from the inconsistency in unit representation and terminology. AI detected formatting inconsistencies, such as non-standard abbreviations and technical jargon that could be confusing for users unfamiliar with the domain. Although this issue was not rated as highly severe, it contributed to a lack of clarity in interpreting results. In contrast, students focused more on the absence of clear input validation, pointing out that the system accepted invalid characters, including letters, without any warning. This led to errors in conversions without clear indications of what went wrong. AI suggested refining the terminology and standardizing the way units were

displayed, while students emphasized the need for stricter validation rules and real-time alerts to notify users of incorrect inputs.

The lack of user control and flexibility further complicated usability. AI noted that the absence of an undo or reset function made it difficult for users to correct mistakes efficiently. This issue was rated as a moderate usability concern, as it forced users to manually clear fields before re-entering their inputs. Students, however, found a more immediate problem: the conversion button itself was not easily noticeable, leading to confusion about how to execute a conversion. They assigned this issue a higher severity level, as users frequently struggled to locate the primary function of the tool. AI recommended integrating a reset function, whereas students suggested improving the prominence and placement of the conversion button to streamline interactions.

Error prevention emerged as another significant usability challenge. AI identified a lack of validation for incorrect inputs, noting that users could enter letters or symbols without receiving immediate feedback. This resulted in errors that were not explained adequately, leading to frustration. AI classified this as a major issue, proposing a stricter validation process and clearer error prompts. Students similarly highlighted this problem, but their concerns were more focused on the phrasing of error messages. Many messages lacked specificity, making it difficult for users to understand what had gone wrong. Their feedback emphasized the need for clearer, more instructional error messages that explicitly guided users toward the correct input format.

In addition to these functional issues, the overall interface design contributed to usability difficulties. AI detected inconsistencies in font sizes, cluttered spacing, and excessive advertisements that disrupted the workflow. The cluttered design was rated as a moderate concern, as it affected readability and increased cognitive load. Students, however, framed this issue differently, noting that the navigation structure was unclear, which made it difficult to locate relevant sections. While AI suggested reducing visual clutter and improving contrast, students focused more on restructuring the layout to ensure that important elements were easily accessible.

A recurring theme across AI and student evaluations was the lack of help and documentation. AI noted that the absence of tool-tips and contextual guidance made it difficult for

users to understand certain features. However, students expressed even stronger concerns about this issue, stating that the website lacked basic instructions on how to use the conversion tool. They rated this as a more severe issue than AI did, as the absence of clear guidance led to confusion and inefficiencies. AI proposed adding tool-tips to explain key functions, while students recommended a dedicated help section with user instructions.

Overall, Engineering Toolbox presented usability issues that were distinct from those in UnitConverters.net. AI's evaluation emphasized technical inconsistencies, such as unclear validation and missing system feedback, while students were more concerned with workflow disruptions, navigation inefficiencies, and the clarity of instructions. Addressing these issues through improved feedback mechanisms, clearer navigation structures, and more intuitive input validation would significantly enhance the user experience.

The comparative severity ratings assigned by AI and students are presented in Table 3.2. Figure 3.4 provides a visual comparison of these assessments, illustrating where AI and human evaluators aligned and where their perspectives differed.

Table 3.2: Comparison of average severity ratings between AI and students evaluations for the *EngineeringToolbox.com* website.

Heuristic	**AI Severity**	**Student Severity**
Visibility of System Status	3.0	2.25
Match Between System and the Real World	2.0	2.75
User Control and Freedom	3.0	2.63
Consistency and Standards	2.0	1.83
Error Prevention	3.75	2.80
Recognition Rather Than Recall	2.0	2.00
Flexibility and Efficiency of Use	2.0	2.56
Aesthetic and Minimalist Design	3.0	2.54
Help Users Recognize, Diagnose, and Recover	3.0	2.00
Help and Documentation	2.0	2.00

3.3.4 Website 3 analysis: Convert me

Compared to the first two websites, where issues primarily revolved around feedback delays and navigation complexity, Convert Me presented a different set of usability challenges, particularly in interface clutter, inconsistent user interactions, and input validation weaknesses. AI and student evaluations both pointed to inefficiencies in how information was presented, difficulties in navigating the site, and confusing input handling. How-

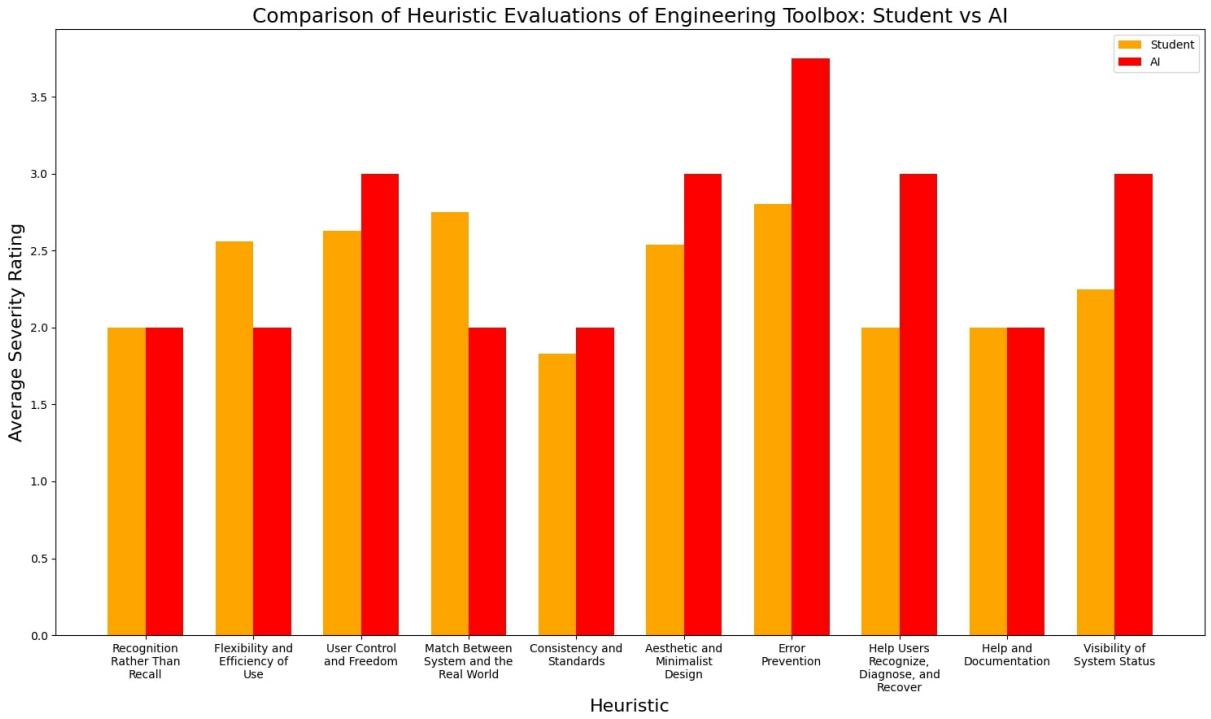

Figure 3.4: Comparison of Heuristic Evaluations for Engineering Toolbox: AI vs. Students

ever, their focus areas differed: AI emphasized structural and functional inconsistencies, while students were more concerned with usability roadblocks that directly affected their workflow.

One of the most pressing concerns for both AI and student evaluators was the disorganized interface, which significantly impacted usability. AI highlighted that the homepage was overloaded with content, the converter tool was hidden under excessive elements and advertisements. This cluttered layout not only made it harder for users to find the conversion tool but also distracted from the primary function of the website. AI assigned a high severity rating to this issue (Severity: 4), recommending a more minimalist design with clearer visual hierarchy. Students expressed similarly this concern, noting that ads often interrupted the conversion process and caused accidental clicks. However, instead of purely reducing clutter, they suggested repositioning the converter to the top of the page for easier accessibility.

The absence of proper input validation and error handling was a significant usability flaw recognized by both AI and students. AI found that the system permitted non-numeric inputs and invalid unit selections without providing immediate feedback, resulting in errors only appearing after users attempted conversions (Severity: 4). It recommended

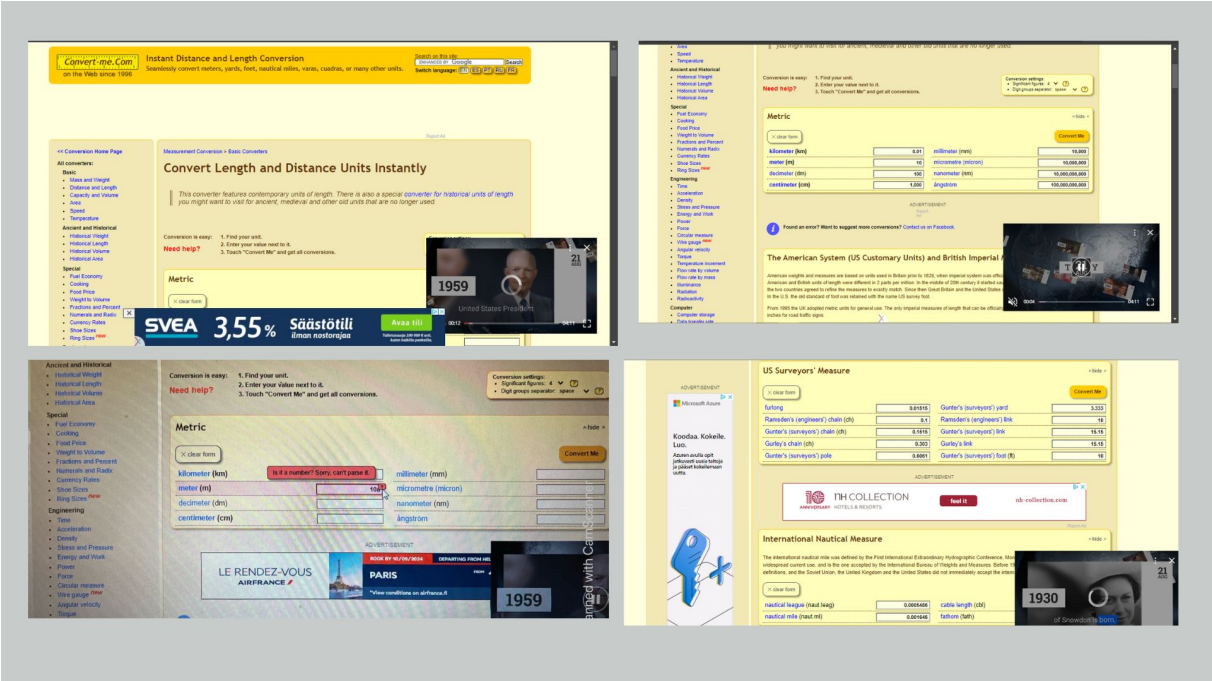

Figure 3.5: Interface of the Convert me.com website (screenshot by the author).

implementing real-time input validation and restricting invalid selections. Students faced similar frustrations but framed the issue differently. They struggled with unclear error messages that provided little guidance on how to correct mistakes. Many of the messages, such as "Is it a number? Sorry, can't parse it," were vague and unhelpful. They rated this issue as a high usability problem and suggested that error messages should be more instructive and placed closer to the affected input fields for better visibility.

The conversion process itself was another area of concern. AI noted that while some conversions happened automatically, a "Convert Me" button was still present, leading to confusion about whether user action was required (Severity: 3). It recommended either removing the redundant button or making it explicitly clear whether conversions were automatic or manual. Students also expressed confusion but were more concerned with the inconsistency in how different converters operated. Some conversion tools redirected them to new pages, while others processed inputs on the same screen, disrupting workflow and making the system unpredictable. They suggested standardizing the behavior across all conversion tools to ensure consistency.

A related problem was the absence of user control options, particularly the lack of an undo/reset function. AI noted that users had no way to clear or revert inputs quickly, forcing them to manually delete values (Severity: 3). It proposed adding a reset button

and undo functionality for convenience. Students, on the other hand, were more focused on how the converter did not remember their last-used units, meaning they had to re-select them every time they returned to the page. They viewed this as an unnecessary usability burden (Severity: 3) and suggested a history or favorites feature to retain frequently used conversions.

Another usability concern was navigation inefficiencies. AI found that the lack of a search function required users to manually scroll through long unit lists to find what they needed (Severity: 2). It recommended introducing a search bar for quicker access. Students had a similar complaint but framed it differently. They noted that certain unit selection menus were hidden or difficult to locate, making the process unnecessarily tedious. Their suggestion was to improve menu placement and visibility rather than just adding a search function.

Lastly, help and documentation deficiencies were noted by both evaluations. AI observed that there were absence of on-screen guidance and contextual hints to guide users, forcing them to figure out conversion processes on their own (Severity: 2). It recommended adding contextual guidance to improve usability. Students found this issue even more frustrating, as help materials were difficult to access and often redirected to external pages, making it inconvenient for users to find relevant information. They rated this as a major usability issue (Severity: 3) and suggested integrating concise, in-page documentation to reduce dependency on external resources.

In summary, Convert Me presented a distinct set of usability challenges compared to the previous websites. While Unit Converters.net and Engineering Toolbox had issues related to feedback delays and navigation inefficiencies, Convert Me struggled with interface clutter, inconsistent interactions, weak error handling, and a lack of user control options. AI's analysis emphasized structural fixes such as layout improvements, better validation, and real-time feedback mechanisms, while students focused more on workflow efficiency, personalization options, and clearer error handling.

To illustrate the differences in AI and student evaluations, Table 3.3 provides a comparative summary of the severity ratings assigned to each heuristic. Figure 3.6 visually presents these findings, highlighting areas of agreement and divergence between the two

evaluation methods.

Table 3.3: Comparison of average severity ratings between AI and students evaluations for the *Convert-me.com* website.

Heuristic	**AI Severity**	**Student Severity**
Aesthetic and Minimalist Design	3.00	2.63
Help and Documentation	2.00	2.50
Error Prevention	3.75	2.13
Visibility of System Status	2.33	2.75
Help Users Recognize, Diagnose, and Recover	2.67	1.80
Recognition Rather Than Recall	2.00	1.50
Consistency and Standards	2.00	3.00
Match Between System and the Real World	1.00	1.50
User Control and Freedom	2.50	2.67
Flexibility and Efficiency of Use	2.50	3.00

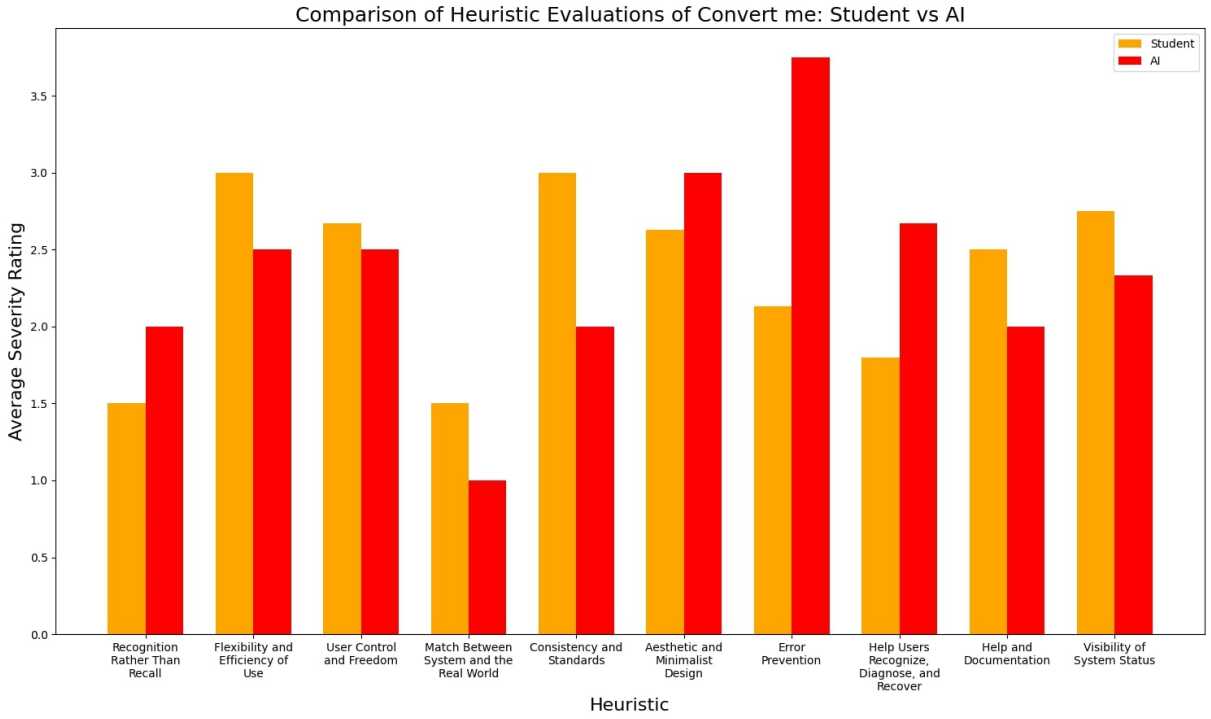

Figure 3.6: Comparison of Evaluations for Convert-Me: AI vs. Students.

3.3.5 Website 4 analysis: Asknumbers

The usability evaluation of AskNumbers uncovered distinct challenges, differing from the previously analyzed websites. While earlier sites suffered from visibility delays and workflow inefficiencies, AskNumbers presented obstacles related to finding essential tools, navigation difficulties, and unclear system feedback. AI's assessment emphasized structural inconsistencies, terminology issues, and validation gaps, whereas students focused

more on practical usability concerns, such as locating the converter, inefficient navigation, and poor error handling.

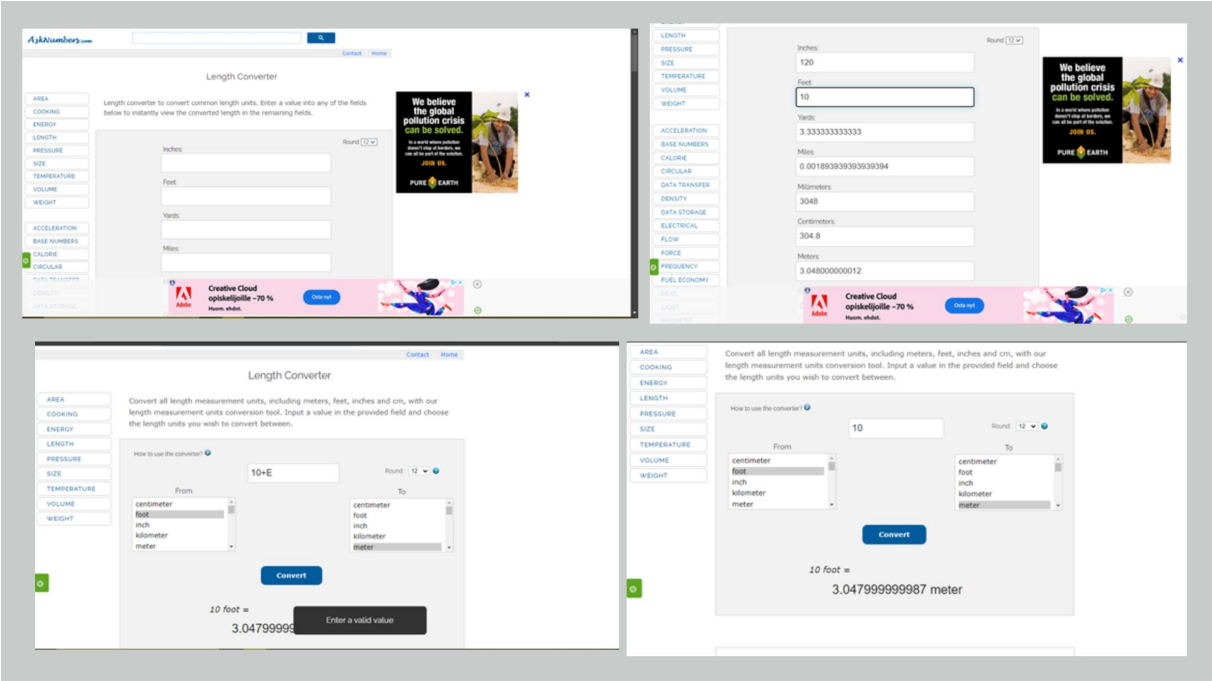

Figure 3.7: Interface of the Asknumbers.com website (screenshot by the author).

A significant usability concern identified in both AI and student evaluations was the difficulty in locating the primary conversion tool. AI highlighted the cluttered layout, noting that excessive elements pushed the key functionality out of immediate view (Severity: 3). Students echoed this concern, stating that the left-aligned layout left excessive empty space on wider screens, making navigation cumbersome (Severity: 3). AI recommended a cleaner layout with a clearer visual hierarchy, while students suggested repositioning the conversion tool for easier access.

The lack of clear system feedback further contributed to usability challenges. AI noted that the website failed to indicate when conversions were in progress or completed, leaving users uncertain if their inputs had been processed (Severity: 2). Students encountered similar problems but framed them differently, emphasizing that some conversions failed silently without warning, particularly when entering large numbers (Severity: 4). AI proposed real-time indicators for process completion, whereas students recommended clearer messaging and visual cues for successful conversions.

Another major concern was input validation and error handling, which both AI and students identified as a major flaw. AI found that the system allowed invalid inputs,

including letters in numeric fields, without providing immediate feedback (Severity: 4). Additionally, there were no warnings for mismatched unit types, which led to incorrect conversions. Students faced similar frustrations but focused on the unclear error messages, which failed to specify what went wrong (Severity: 3). AI suggested implementing stricter validation and real-time feedback, while students emphasized the need for clearer, more instructive error messages.

The lack of user control options further complicated the usability experience. AI identified the absence of an undo/reset function, which forced users to manually clear inputs instead of quickly resetting the form (Severity: 3). Additionally, the inability to switch between units without re-entering values made the process inefficient. Students shared these concerns but were more frustrated with the website requiring form resubmission when navigating back, which interrupted the workflow (Severity: 3). AI recommended adding an undo/reset button, while students suggested streamlining navigation to prevent unnecessary page reloads.

Another critical issue was navigation inefficiencies, which disrupted the usability experience. AI pointed out that the website lacked a search function for locating unit conversions, requiring users to manually scroll through long lists (Severity: 2). Students encountered similar issues but noted that menus were poorly placed and difficult to locate, making the process unnecessarily tedious (Severity: 3). AI suggested adding a search bar, whereas students recommended reorganizing menu structures to ensure key features were easily accessible.

In terms of consistency and standards, both AI and students identified discrepancies across different conversion tools. AI noted terminology inconsistencies, such as variations in unit names and drop-down styles (Severity: 2). Students encountered a more immediate issue—different converters operated in different ways, with some requiring new page loads while others processed conversions in place (Severity: 3). They also flagged the currency converter as problematic, as it often failed to work correctly without explanation (Severity: 4). AI recommended standardizing terminology and interface elements, while students suggested ensuring uniform behavior across all conversion tools.

Lastly, help and documentation deficiencies were a shared concern. AI noted the

absence of in-page tool-tips and contextual guidance, which forced users to figure out conversion processes on their own (Severity: 2). Students found this issue more frustrating, particularly when help materials redirected them to external pages, disrupting their experience (Severity: 3). AI suggested integrating contextual tool-tips, while students recommended providing on-page help to reduce reliance on external resources.

The AskNumbers evaluation revealed a set of usability challenges distinct from the previously analyzed websites. Locating the main conversion tool, poor navigation, unclear system feedback, and weak error handling stood out as the most critical issues. While AI emphasized structural inconsistencies and missing validation mechanisms, students focused more on workflow disruptions and interface usability barriers. Addressing these concerns through a cleaner layout, real-time validation, standardized interactions, and clearer messaging would significantly enhance the overall user experience.

The comparative severity ratings assigned by AI and students are presented in Table 3.4, which summarizes key usability concerns. Figure 3.8 provides a visual representation of these assessments, highlighting areas of agreement and divergence between the two evaluation methods.

Table 3.4: Comparison of average severity ratings between AI and students evaluations for the *Asknumbers.com* website.

Heuristic	**AI Severity**	**Student Severity**
Aesthetic and Minimalist Design	2.50	1.89
Help and Documentation	2.00	3.00
Error Prevention	3.67	3.00
Visibility of System Status	1.33	2.60
Help Users Recognize, Diagnose, and Recover	3.50	2.80
Recognition Rather Than Recall	2.00	2.00
Consistency and Standards	1.67	2.00
User Control and Freedom	3.00	2.50
Flexibility and Efficiency of Use	1.67	2.67
Match Between System and the Real World	2.67	-

3.4 Comparative analysis across all websites

The usability evaluation of the four websites revealed distinct differences in how AI and students identified and prioritized usability issues. AI predominantly detected technical

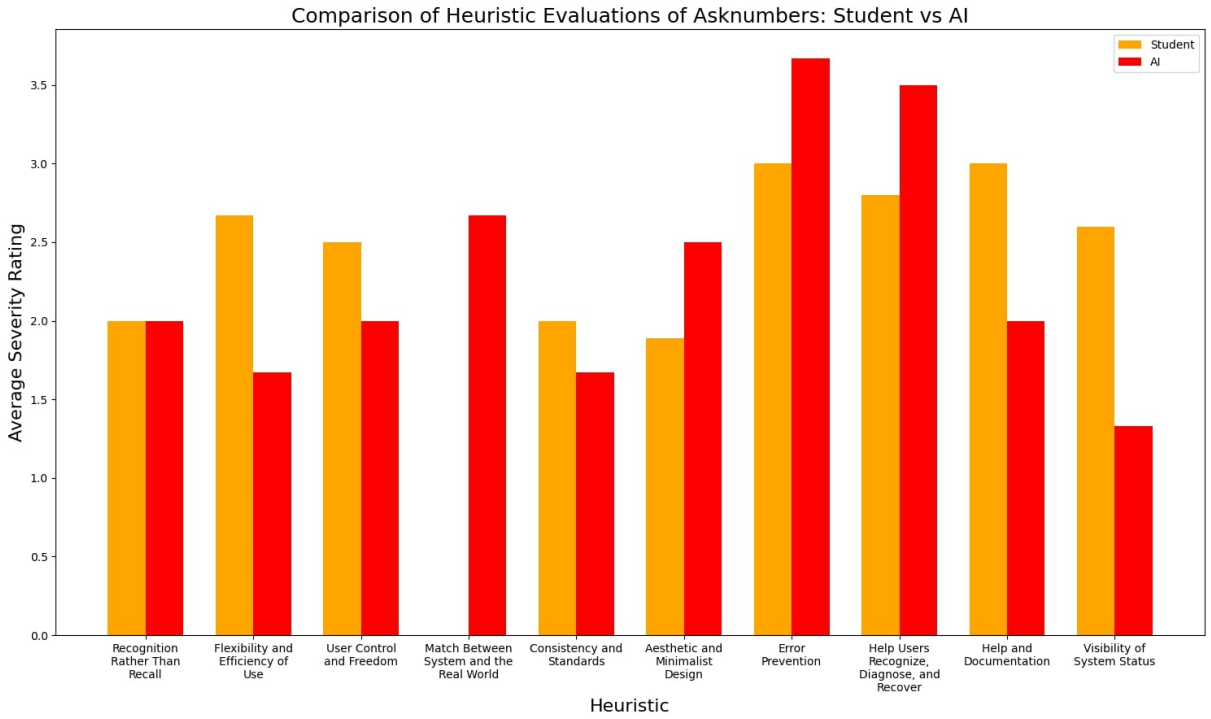

Figure 3.8: Comparison of Evaluations for AskNumbers: AI vs. Students.

inconsistencies, error handling flaws, and deviations from standard usability guidelines, while students focused more on practical user experience barriers, such as clarity of instructions, navigation efficiency, and interaction flexibility. This divergence in evaluation perspectives highlights the complementary strengths of AI-driven and human assessments in usability testing, as also emphasized by Fan et al. (2022), who found that combining AI explanations with human judgment improved the overall effectiveness of UX evaluations [65].

3.4.1 Tool ranking by AI and students

The rankings of the best and worst tools, as determined by student preferences, are illustrated in Figure 3.9. These rankings provide insights into how usability factors influenced user perception.

Student rankings

Students rated the tools based on interface simplicity, ease of use, and clarity of feedback. As shown in the left pie chart of Figure 3.9, the majority (56.6%) considered UnitConverters.net the most user-friendly tool, citing its clear layout, minimal distractions, and

easily accessible features. AskNumbers followed at 21.7%, though concerns about unclear error messages and confusing terminology were frequently mentioned.

On the other hand, the right pie chart reveals that Convert-Me was the least preferred tool (40%), with students highlighting its excessive advertisements and cluttered interface as major usability barriers. Engineering Toolbox also received negative feedback (30%), primarily due to its outdated design, complex navigation, and lack of system feedback.

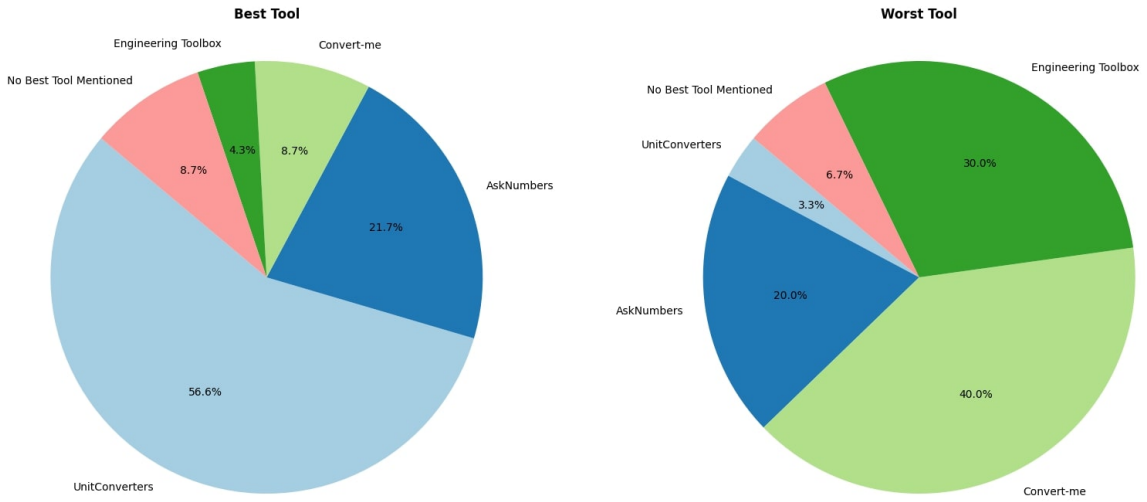

Figure 3.9: Students' Rankings of Best and Worst Tools.

AI rankings

AI rankings followed a different pattern, prioritizing technical consistency, structured layouts, and error prevention mechanisms. Like students, AI identified UnitConverters.net as the best tool, but its reasoning differed. It was ranked highly due to strong input validation, structured layout, and effective system feedback. AskNumbers was also rated second, primarily for its automated conversion process, though AI flagged issues such as weak error handling and inconsistent UI elements. Conversely, Convert-Me and Engineering Toolbox received the lowest rankings, with AI pinpointing unstructured interfaces, inconsistent UI behaviors, and inadequate feedback systems as key drawbacks.

3.4.2 Usability patterns across heuristics

Figure 3.10 provides a heatmap representation of severity ratings assigned by AI and students across the ten heuristic principles. Darker shades indicate higher severity ratings,

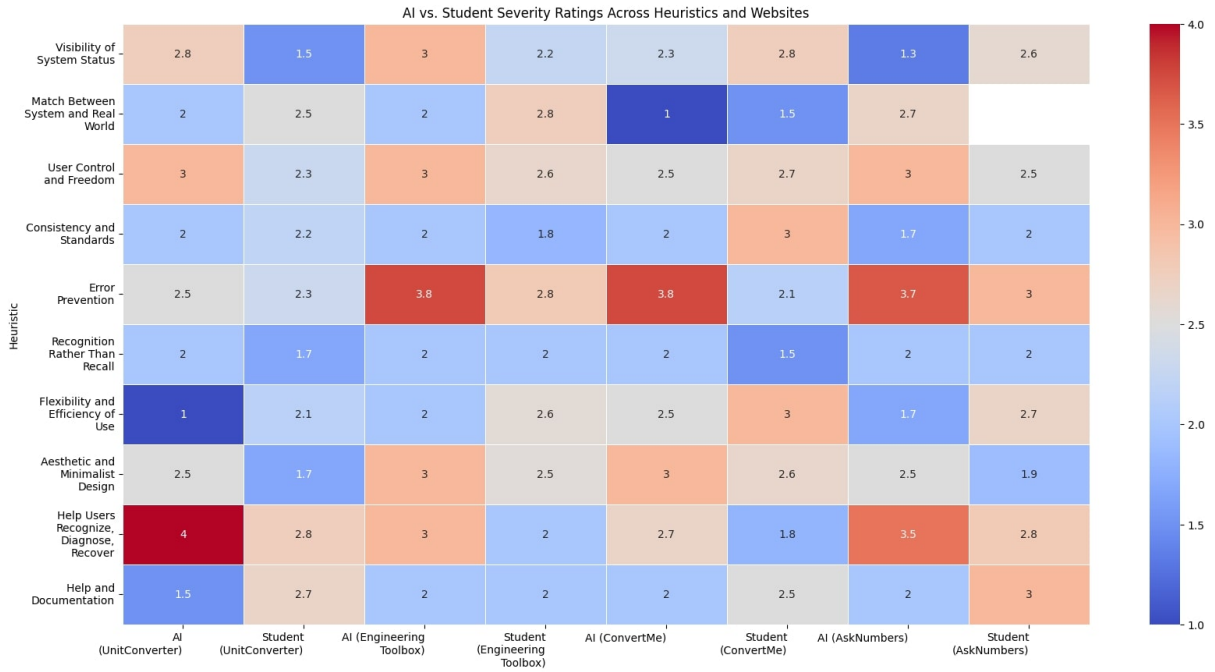

Figure 3.10: Heatmap Comparing AI and Student Severity Ratings Across Heuristics.

highlighting key usability concerns that emerged across multiple websites.

A comparative analysis of heuristic evaluations reveals recurring patterns in usability challenges. Table 3.5 summarizes the most prominent issues identified by AI and students, along with the websites most affected by these concerns.

The heatmap and table reveal that Convert-Me, AskNumbers, and UnitConverters.net had the most critical flaws in error prevention, with AI highlighting weak input validation and vague error messages, while students struggled with poor system feedback and unclear error handling. Engineering Toolbox and UnitConverters.net were noted for feedback mechanism issues, where AI detected the absence of progress indicators, and students reported that missing real-time responses created confusion.

Similarly, Convert-Me and Engineering Toolbox were flagged for cluttered interfaces, with AI noting inconsistent layouts, while students found excessive advertisements distracting. In terms of help and documentation, students were particularly critical of Engineering Toolbox, AskNumbers, and UnitConverters.net, stating that instructions were difficult to access or missing altogether, whereas AI assigned a lower severity rating to this category, as its focus was more on interface consistency rather than user guidance.

In this study, AI excelled at identifying structural inconsistencies and adherence to heuristics. It did not fully capture the human experience of usability challenges. Students,

Table 3.5: Comparison of usability issues identified by AI and students.

Issue	**AI identified**	**Student identified**	**Most affected websites**
Error prevention	Missing input validation, no automated checks, vague error messages	Confusing error messages, incorrect input handling	Convert-Me, AskNumbers, UnitConverters.net, Engineering Toolbox
Feedback mechanisms	No progress indicators, unclear system status, lack of real-time validation	Delayed feedback, missing confirmation messages	Engineering Toolbox, AskNumbers, UnitConverters.net, Convert-Me
Aesthetic and minimalist design	Inconsistent layout, poor spacing, excessive elements	Visual distractions, overwhelming UI, too many ads	Convert-Me, Engineering Toolbox
Help and documentation	Low severity rating, missing tooltips, lack of contextual guidance	Missing/unclear instructions, no search function	AskNumbers, Engineering Toolbox, UnitConverters.net
Navigation and flexibility	No search function, difficult menu structure, inefficient dropdowns	Hard-to-find features, confusing navigation	Convert-Me, Engineering Toolbox, UnitConverters.net, AskNumbers
User control and freedom	No undo/reset function, limited flexibility, difficult back-navigation	Frustration with non-reversible actions, inability to swap units	AskNumbers, Convert-Me, UnitConverters.net, Engineering Toolbox
Consistency and standards	Different UI behavior across conversion tools, non-uniform design	Unclear UI interactions, inconsistent unit formatting	AskNumbers, Convert-Me, UnitConverters.net
Help users recognize, diagnose, and recover from errors	Unclear error messages, weak troubleshooting guidance	Hard-to-interpret error prompts, lack of guidance on fixing errors	AskNumbers, Engineering Toolbox, Convert-Me
Match between system and the real world	Confusing terminology, unit labels not intuitive	Unfamiliar language, difficult-to-interpret conversion results	UnitConverters.net, AskNumbers

on the other hand, highlighted interaction barriers, instructional clarity, and real-world usability issues that AI tends to overlook. These differences will be further examined in the discussion chapter.

4 Discussion

4.1 AI in usability testing: strengths and limitations

The study's results indicate that AI can effectively and swiftly detect structural usability problems, particularly those involving input validation, consistent interface design, and feedback from the system. Across all four websites, AI and human evaluators identified many of the same usability flaws, reinforcing AI's reliability in heuristic-based assessments. This alignment suggests that AI has the potential to assist in automating usability testing, which may help reduce the time and effort required for large-scale evaluations.

This study may not give a full picture of AI's capabilities for usability testing. AI was constrained by only using static screenshots as input. AI was thus unable to assess interactive elements such as hover effects, progress indicators, and real-time feedback mechanisms. A notable example was AI repeatedly flagging missing system status indicators, even when they appeared dynamically during live interactions. This limitation highlights the need for AI models that can interact with interfaces in real time, rather than relying solely on static image analysis.

While AI excels in detecting rule-based inconsistencies, human evaluators remain essential for validating its findings and assessing dynamic usability aspects. In this study, AI seemed to apply heuristic principles consistently, which sometimes led to repeated detection of similar issues across different websites. While this ensured consistency, it also meant that AI sometimes flagged issues without considering their context. For example, AI identified cluttered layouts across multiple websites but failed to distinguish between clutter caused by excessive advertisements and clutter resulting from poor content organization. In contrast, human evaluators provided qualitative insights into how such design elements impacted user experience. This distinction underscores AI's lack of contextual

awareness; while it can identify usability violations, it does not interpret their effect on user engagement and cognitive load.

4.2 Comparative analysis of AI and human evaluations

A key observation from this study is that AI and human evaluators identified several overlapping usability issues, suggesting that AI may be useful in detecting rule-based inconsistencies. This finding is crucial in addressing how AI can complement human abilities to create a more efficient and accurate usability testing process. By automating routine tasks, AI greatly minimizes the time and effort needed for usability testing, enabling human evaluators to concentrate on more subjective elements like cognitive load and emotional engagement.

However, AI's evaluations lacked the depth of human assessments, particularly in areas involving user perception and interaction difficulties. Human evaluators focused on how usability flaws affected workflow, cognitive effort, and emotional responses. For example, AI accurately identified overcrowded layouts in Convert-Me, but it did not recognize how intrusive advertisements disrupted user workflow. Although GPT-4o extracted text from the interface screenshot, it lacked contextual understanding of which elements were essential to the user task and which were peripheral or disruptive, such as ads. Human participants, on the other hand, noted how these distractions increased task completion time and user frustration. Similarly, AI flagged inconsistent formatting in Engineering Toolbox but failed to account for how poor result visibility affected user comprehension. These examples illustrate AI's inability to assess the human experience of usability beyond surface level inconsistencies.

The findings indicate that AI appears to be particularly useful for automating large-scale usability evaluations, potentially allowing human testers to focus more on the nuanced aspects of user experience. Rather than viewing AI as a replacement for human evaluators, it should be seen as a tool that complements human expertise, ensuring both efficiency and depth in usability testing.

4.3 Methodological insights and query optimization

Another important observation was that AI did not identify all usability issues in a single evaluation. Instead, multiple queries were required to capture a more comprehensive set of findings. The breadth of AI-generated results was influenced by how queries were framed—when prompts were refined to focus on specific usability heuristics, AI identified additional issues that were initially overlooked.

This iterative nature of AI evaluation suggests that structured, multi-stage querying can improve the accuracy and depth of AI-generated findings. Rather than relying on a single assessment, usability testers may enhance AI's output by rephrasing prompts and requesting issue breakdowns by heuristic category. This strategy not only maximizes AI's detection capabilities but also ensures that its findings align more closely with human evaluations.

4.4 Scope and applicability of AI in usability testing

This study focused exclusively on website usability testing to keep the scope manageable and allow for a clear comparison between AI and human evaluations. Websites were chosen as the test case because they present fundamental usability challenges, such as navigation, form validation, and feedback mechanisms, which AI could analyze based on heuristic principles. However, the possible uses of AI in usability testing extend far beyond website evaluations.

Artificial intelligence can support usability testing across multiple fields, such as mobile applications, enterprise software, and embedded systems. In software testing, AI offers the capability to automate various aspects of functional testing, detect usability bottlenecks, and analyze user interaction patterns on a much larger scale. However, these applications were beyond the scope of this study, which aimed to determine whether AI could be effectively used in usability testing at a fundamental level. Future research should explore AI's role in more complex usability scenarios, including real-time software interaction testing and AI-driven user behavior analysis.

4.5 Scalability and practical applications

One of AI's key advantages in usability testing is its scalability. Unlike human testers, who require coordination, training, and time investment, AI can execute repeated assessments with minimal resource allocation. This suggests that AI could be particularly useful for early-stage usability testing, where developers need rapid feedback before conducting more in depth human evaluations. Additionally, AI-generated usability reports could serve as learning tools for novice developers, helping them identify fundamental design flaws without requiring specialized usability expertise.

However, while AI reduces manual workload, it does not eliminate the need for human involvement. Its inability to assess dynamic elements, contextual usability challenges, and subjective user experiences means that AI-generated reports should be validated through human review. The most effective approach is to integrate AI as an initial diagnostic tool, using its efficiency to highlight potential usability concerns while relying on human testers to interpret and refine its findings.

4.6 Future Directions and Enhancements

A key limitation of this study was AI's reliance on static images. Future research should explore AI models capable of analyzing video recordings of user interactions. Video based evaluation could provide a more comprehensive understanding of usability by capturing sequential interactions, user navigation patterns, and real time system responses. By incorporating temporal data, AI could more accurately assess system status visibility, task completion flows, and responsiveness issues that are necessarily missed in static image based analysis.

An even more promising direction would be the development of AI models that can actively interact with websites, mimicking real user behavior to uncover usability issues. AI-driven autonomous interaction, including navigating through pages, filling out forms, and simulating user tasks, has the potential to offer deeper insights into navigation efficiency, input validation failures, and real-time feedback accuracy. If AI could perform structured usability tests by interacting with different interface components, it would

allow for a far more realistic and adaptable evaluation process.

Additionally, future advancements should integrate real time AI-powered usability testing frameworks that combine interaction tracking with machine learning based usability predictions. Such systems could detect usability bottlenecks dynamically, adapting their evaluation based on user interaction data rather than relying solely on predefined heuristic principles. AI models that incorporate eye-tracking, mouse movement heatmaps, and behavioral analysis could further enhance usability assessments, making them more responsive to real-world user experiences.

Expanding AI's capabilities beyond static heuristic analysis would significantly improve its role in usability testing. By enabling real-time interaction with digital interfaces, AI could evolve from a rule based evaluator into an adaptive usability testing tool that closely mirrors human testing methodologies. Further research is needed to unify these developments into a comprehensive, adaptable, and interactive AI-driven usability testing framework.

5 Conclusion

This study focused on understanding the impact of artificial intelligence (AI) in usability testing, highlighting its strengths and limitations compared to human evaluations. The findings suggest that the specific AI model used—GPT-4o—was effective in detecting structural usability issues, input validation errors, and interface inconsistencies with speed and consistency. However, it lacked the ability to fully understand user context, dynamic interactions, and emotional responses, making human evaluators indispensable in capturing the experiential and subjective aspects of usability.

The study suggest that AI's scalability makes it a powerful tool for early-stage usability testing and large-scale audits. It hold promise for reducing costs and effort by automating repetitive evaluations, allowing human testers to focus on subjective and experiential aspects of usability. However, AI's reliance on static images limits its ability to assess real-time interactions, requiring further advancements to enhance its adaptability to dynamic user experiences.

The study underscores the significance of integrating both methods, where AI serves as an efficient, data-driven evaluator, while human testers provide critical contextual insights. This combination ensures a more comprehensive usability evaluation process that balances automation with human expertise. Further studies should prioritize the creation of AI models with capabilities for live assessments, behavioral analysis, and dynamic learning to strengthen both accuracy and contextual relevance in usability evaluations.

In conclusion, AI can be a valuable asset in usability testing, streamlining the evaluation process and reducing costs. However, human oversight remains essential to interpret findings meaningfully and ensure usability testing remains user-centric. By integrating AI's analytical strengths with human insight, usability testing can evolve into a more efficient, adaptive, and insightful process, aligning with the rapid advancements in digital

technology.

References

[1] J. Rubin and D. Chisnell, *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, 2nd. Indianapolis, IN: Wiley Publishing, Inc, 2008, Foreword by Jared Spool, ISBN: 978-0-470-18548-3.

[2] C. Barnum, *Usability Testing Essentials: Ready, Set ... Test!* San Francisco: Morgan Kaufmann, 2010, ISBN: 978-0-12-375092-1. DOI: 10.1016/C2009-0-20478-8. [Online]. Available: https://www.sciencedirect.com/book/9780123750921/usability-testing-essentials?via=ihub.

[3] J. Nielsen, *Report from a 1994 web usability study*, Published on December 18, 1994, 1994. [Online]. Available: https://www.nngroup.com/articles/1994-web-usability-report/.

[4] S. Krug, *Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems* (Voices That Matter), 1st. Berkeley, CA: New Riders, 2009, ISBN: 978-0321657299.

[5] M. Al-Razgan, R. Aldossari, L. Albesher, *et al.*, "Challenges of integrating agile and ux/ucd: Systematic literature review", *International Journal of Computer Applications*, vol. 184, no. 33, pp. 40–58, Oct. 2022, ISSN: 0975-8887. DOI: 10.5120/ijca2022922426. [Online]. Available: https://ijcaonline.org/archives/volume184/number33/32529-2022922426/.

[6] B. Fu and B. Steichen, "Using behavior data to predict user success in ontology class mapping - an application of machine learning in interaction analysis", in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 216–223. DOI: 10.1109/ICOSC.2019.8665670. [Online]. Available: https://ieeexplore.ieee.org/document/8665670.

[7] S. Amershi and C. Conati, "Combining unsupervised and supervised classification to build user models for exploratory learning environments", *Journal of Educational Data Mining*, Sep. 2009. [Online]. Available: https://www.microsoft.com/en-us/research/publication/combining-unsupervised-supervised-classification-build-user-models-exploratory-learning-environments/.

[8] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study", *Applied Sciences*, vol. 11, no. 9, 2021, ISSN: 2076-3417. DOI: 10.3390/app11093986. [Online]. Available: https://www.mdpi.com/2076-3417/11/9/3986.

[9] D. Zhou, "Human-computer interaction interface design in intelligent medical system under the background of artificial intelligence", in *2022 International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*, 2022, pp. 10–13. DOI: 10.1109/AIARS57204.2022.00010.

[10] D. Sarma and M. K. Bhuyan, "Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review", *SN Computer Science*, vol. 2, p. 436, 2021. DOI: 10.1007/s42979-021-00827-x. [Online]. Available: https://link.springer.com/article/10.1007/s42979-021-00827-x.

[11] Y. Shi, "A bibliometric analysis of eye tracking in user experience research", in *Human-Computer Interaction. HCII 2024*, ser. Lecture Notes in Computer Science, M. Kurosu and A. Hashizume, Eds., vol. 14684, Springer, Cham, 2024, ISBN: 978-3-031-60404-1. DOI: 10.1007/978-3-031-60405-8_12. [Online]. Available: https://doi.org/10.1007/978-3-031-60405-8_12.

[12] D. Drungilas, I. Ramašauskas, and M. Kurmis, "Emotion recognition in usability testing: A framework for improving web application ui design", *Applied Sciences*, vol. 14, no. 11, p. 4773, 2024. DOI: 10.3390/app14114773. [Online]. Available: https://www.mdpi.com/2076-3417/14/11/4773.

[13] B. Cabrero-Daniel, "AI for Agile Development: A Meta-Analysis", *arXiv preprint arXiv:2305.08093*, 2023, Preprint, University of Gothenburg. DOI: 10.48550/arXiv.2305.08093. [Online]. Available: https://arxiv.org/abs/2305.08093.

[14] A. BAHI, J. GHARIB, and Y. GAHI, "Integrating generative ai for advancing agile software development and mitigating project management challenges", *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, 2024. DOI: 10.14569/IJACSA.2024.0150306. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2024.0150306.

[15] F. A. Batarseh and A. J. Gonzalez, "Predicting failures in agile software development through data analytics", *Software Quality Journal*, vol. 26, pp. 49–66, 2018. DOI: 10.1007/s11219-015-9285-3. [Online]. Available: https://doi.org/10.1007/s11219-015-9285-3.

[16] A. Namoun, A. Alrehaili, Z. U. Nisa, H. Almoamari, and A. Tufail, "Predicting the usability of mobile applications using ai tools: The rise of large user interface models, opportunities, and challenges", *Procedia Computer Science*, vol. 238, pp. 671–682, 2024, The 15th International Conference on Ambient Systems, Networks and Technologies Networks (ANT) / The 7th International Conference on Emerging Data and Industry 4.0 (EDI40), April 23-25, 2024, Hasselt University, Belgium, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2024.06.076. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924013127.

[17] N. van Berkel, M. B. Skov, and J. Kjeldskov, "Human-ai interaction: Intermittent, continuous, and proactive", *Interactions*, vol. 28, no. 6, pp. 67–71, Nov. 2021, ISSN: 1072-5520. DOI: 10.1145/3486941. [Online]. Available: https://doi.org/10.1145/3486941.

[18] J. C. Bastien, "Usability testing: A review of some methodological and technical aspects of the method", *International Journal of Medical Informatics*, vol. 79, no. 4, e18–e23, 2010, Human Factors Engineering for Healthcare Applications Special Issue, ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2008.12.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505608002098.

[19] L. M. Amugongo, A. Kriebitz, A. Boch, *et al.*, "Operationalising ai ethics through the agile software development lifecycle: A case study of ai-enabled mobile health applications", *AI Ethics*, 2023, Published 15 August 2023, Accepted 02 August 2023,

Received 31 May 2023. DOI: 10.1007/s43681-023-00331-3. [Online]. Available: https://doi.org/10.1007/s43681-023-00331-3.

[20] M. Asghar, I. S. Bajwa, S. Ramzan, H. Afreen, S. Abdullah, and M. Kumar, "A genetic algorithm-based support vector machine approach for intelligent usability assessment of m-learning applications", *Mobile Information Systems*, vol. 2022, Jan. 2022, ISSN: 1574-017X. DOI: 10.1155/2022/1609757. [Online]. Available: https://doi.org/10.1155/2022/1609757.

[21] J. Hussain, A. U. Hassan, H. S. M. Bilal, *et al.*, "Model-based adaptive user interface based on context and user experience evaluation", *Journal on Multimodal User Interfaces*, vol. 12, pp. 1–16, 2018, Received: 11 December 2016; Accepted: 15 January 2018; Published: 01 February 2018; Issue Date: March 2018. DOI: 10.1007/s12193-018-0258-2. [Online]. Available: https://doi.org/10.1007/s12193-018-0258-2.

[22] A. Khamaj and A. M. Ali, "Adapting user experience with reinforcement learning: Personalizing interfaces based on user behavior analysis in real-time", *Alexandria Engineering Journal*, vol. 95, pp. 164–173, 2024, ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2024.03.045. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016824002874.

[23] S. Roychowdhury, E. Alareqi, and W. Li, "Opam: Online purchasing-behavior analysis using machine learning", in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533658.

[24] G. Martín, A. Fernández-Isabel, I. M. de Diego, *et al.*, "A survey for user behavior analysis based on machine learning techniques: Current models and applications", *Applied Intelligence*, vol. 51, pp. 6029–6055, 2021, Accepted: 16 December 2020; Published: 26 January 2021; Issue Date: August 2021. DOI: 10.1007/s10489-020-02160-x. [Online]. Available: https://doi.org/10.1007/s10489-020-02160-x.

[25] N. Maslej, L. Fattorini, R. Perrault, *et al.*, "Artificial intelligence index report 2024", May 2024, Submitted on 29 May 2024. DOI: 10.48550/arXiv.2405.19522. arXiv: 2405.19522 [cs.AI]. [Online]. Available: https://doi.org/10.48550/arXiv.2405.19522.

[26] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study", *Applied Sciences*, vol. 11, no. 9, p. 3986, 2021. DOI: 10.3390/app11093986. [Online]. Available: https://doi.org/10.3390/app11093986.

[27] S. Borsci, A. Malizia, M. Schmettow, *et al.*, "The chatbot usability scale: The design and pilot of a usability scale for interaction with ai-based conversational agents", *Personal and Ubiquitous Computing*, vol. 26, pp. 95–119, 2022, Published February 2022, Accepted 29 May 2021, Received 22 November 2020. DOI: 10.1007/s00779-021-01582-9. [Online]. Available: https://doi.org/10.1007/s00779-021-01582-9.

[28] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey", *Information Systems*, vol. 112, p. 102 131, 2023, ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2022.102131. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437922001090.

[29] M. Adam, M. Wessel, and A. Benlian, "Ai-based chatbots in customer service and their effects on user compliance", *Electronic Markets*, vol. 31, no. 2, pp. 427–445, 2021, Received: 18 July 2019; Accepted: 12 February 2020; Published: 17 March 2020; Issue Date: June 2021. DOI: 10.1007/s12525-020-00414-7. [Online]. Available: https://doi.org/10.1007/s12525-020-00414-7.

[30] J. T. Liang, C. Yang, and B. A. Myers, "A large-scale survey on the usability of ai programming assistants: Successes and challenges", in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ser. ICSE '24, Lisbon, Portugal: Association for Computing Machinery, 2024, ISBN: 9798400702174. DOI: 10.1145/3597503.3608128. [Online]. Available: https://doi.org/10.1145/3597503.3608128.

[31] S. Sharma, P. Verma, R. Singh, and K. Tripathi, "Advancements in facial expression recognition: A comprehensive analysis of techniques", in *Machine Learning, Image Processing, Network Security and Data Sciences. MIND 2023*, ser. Communications in Computer and Information Science, N. Chauhan, D. Yadav, G. Verma, B. Soni,

and J. Lara, Eds., vol. 2128, Springer, Cham, 2024, ISBN: 978-3-031-62216-8. DOI: 10.1007/978-3-031-62217-5_18. [Online]. Available: https://doi.org/10.1007/978-3-031-62217-5_18.

[32] M. Mohana and P. Subashini, "Facial expression recognition using machine learning and deep learning techniques: A systematic review", *SN Computer Science*, vol. 5, p. 432, 2024. DOI: 10.1007/s42979-024-02792-7. [Online]. Available: https://doi.org/10.1007/s42979-024-02792-7.

[33] L. Ball and B. Richardson, "Eye movement in user experience and human–computer interaction research", in *Eye Tracking*, ser. Neuromethods, S. Stuart, Ed., vol. 183, Humana, New York, NY, 2022, ISBN: 978-1-0716-2390-9. DOI: 10.1007/978-1-0716-2391-6_10. [Online]. Available: https://doi.org/10.1007/978-1-0716-2391-6_10.

[34] H. Li, M. Yazdi, A. Nedjati, *et al.*, "Harnessing ai for project risk management: A paradigm shift", in Springer, Mar. 2024, ISBN: 978-3-031-51718-1. DOI: 10.1007/978-3-031-51719-8_16.

[35] E. Khanna, R. Popli, and N. Chauhan, "Artificial intelligence based risk management framework for distributed agile software development", in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2021, pp. 657–660. DOI: 10.1109/SPIN52536.2021.9566000.

[36] M. Yazdi, E. Zarei, S. Adumene, and A. Beheshti, "Navigating the power of artificial intelligence in risk management: A comparative analysis", *Safety*, vol. 10, no. 2, 2024, ISSN: 2313-576X. DOI: 10.3390/safety10020042. [Online]. Available: https://www.mdpi.com/2313-576X/10/2/42.

[37] Z. Zong and Y. Guan, "Ai-driven intelligent data analytics and predictive analysis in industry 4.0: Transforming knowledge, innovation, and efficiency", *Journal of the Knowledge Economy*, vol. 16, pp. 864–903, 2025. DOI: 10.1007/s13132-024-02001-z. [Online]. Available: https://doi.org/10.1007/s13132-024-02001-z.

[38] S. Kalogiannidis, D. Kalfas, O. Papaevangelou, G. Giannarakis, and F. Chatzitheodoridis, "The role of artificial intelligence technology in predictive

risk assessment for business continuity: A case study of greece", *Risks*, vol. 12, no. 2, 2024, ISSN: 2227-9091. DOI: 10.3390/risks12020019. [Online]. Available: https://www.mdpi.com/2227-9091/12/2/19.

[39] E. Siegel, *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley, 2016. [Online]. Available: https://www.predictiveanalyticsworld.com/book/pdf/Predictive_Analytics_by_Eric_Siegel_Excerpts.pdf.

[40] H. Kang and C. Lou, "Ai agency vs. human agency: Understanding human–ai interactions on tiktok and their implications for user engagement", *Journal of Computer-Mediated Communication*, vol. 27, no. 5, zmac014, Sep. 2022, Published: 18 August 2022, Received: 04 December 2021, Revision received: 23 June 2022, Accepted: 15 July 2022. DOI: 10.1093/jcmc/zmac014. [Online]. Available: https://doi.org/10.1093/jcmc/zmac014.

[41] K. E. S. Souza, M. C. R. Seruffo, H. D. De Mello, D. D. S. Souza, and M. M. B. R. Vellasco, "User experience evaluation using mouse tracking and artificial intelligence", *IEEE Access*, vol. 7, pp. 96 506–96 515, 2019. DOI: 10.1109/ACCESS.2019.2927860.

[42] M. Çolak, İ. Kaya, A. Karaşan, *et al.*, "Two-phase multi-expert knowledge approach by using fuzzy clustering and rule-based system for technology evaluation of unmanned aerial vehicles", *Neural Computing and Applications*, vol. 34, pp. 5479–5495, 2022. DOI: 10.1007/s00521-021-06694-0. [Online]. Available: https://doi.org/10.1007/s00521-021-06694-0.

[43] A. Boodaghian Asl, J. Raghothama, A. Darwich, *et al.*, "A hybrid modeling approach to simulate complex systems and classify behaviors", *Netw Model Anal Health Inform Bioinforma*, vol. 13, no. 9, 2024, Published 18 March 2024, Accepted 07 February 2024, Revised 16 January 2024, Received 12 September 2023. DOI: 10.1007/s13721-024-00446-5. [Online]. Available: https://doi.org/10.1007/s13721-024-00446-5.

[44] X. Jin, H. Dong, M. Evans, and A. Yao, "Design heuristics for artificial intelligence: Inspirational design stimuli for supporting ux designers in generating ai-powered

ideas", in *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts)*, Yokohama, Japan: Association for Computing Machinery, 2021, pp. 1–8, ISBN: 9781450380959. DOI: 10.1145/3411763.3451727. [Online]. Available: https://doi.org/10.1145/3411763.3451727.

[45] S.-C. Necula and V.-D. Păvăloaia, "Ai-driven recommendations: A systematic review of the state of the art in e-commerce", *Applied Sciences*, vol. 13, no. 9, 2023, ISSN: 2076-3417. DOI: 10.3390/app13095531. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5531.

[46] F. Messaoudi and M. Loukili, "E-commerce personalized recommendations: A deep neural collaborative filtering approach", *Operations Research Forum*, vol. 5, no. 5, 2024. DOI: 10.1007/s43069-023-00286-5. [Online]. Available: https://doi.org/10.1007/s43069-023-00286-5.

[47] D. Cirqueira, D. Nedbal, M. Helfert, and M. Bezbradica, "Scenario-based requirements elicitation for user-centric explainable ai", in *Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science, vol 12279*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds., Springer, Cham, 2020. DOI: 10.1007/978-3-030-57321-8_18. [Online]. Available: https://doi.org/10.1007/978-3-030-57321-8_18.

[48] V. D. Karalis, "The integration of artificial intelligence into clinical practice", *Applied Biosciences*, vol. 3, no. 1, pp. 14–44, 2024, ISSN: 2813-0464. DOI: 10.3390/applbiosci3010002. [Online]. Available: https://www.mdpi.com/2813-0464/3/1/2.

[49] P. Padmasiri, P. Kalutharage, N. Jayawardhane, and J. Wickramarathne, "Ai-driven user experience design: Exploring innovations and challenges in delivering tailored user experiences", in *2023 8th International Conference on Information Technology Research (ICITR)*, 2023, pp. 1–6. DOI: 10.1109/ICITR61062.2023.10382802.

[50] K. S. Kaswan, J. S. Dhatterwal, K. Malik, and A. Baliyan, "Generative ai: A review on models and applications", in *2023 International Conference on Communication,*

Security and Artificial Intelligence (ICCSAI), 2023, pp. 699–704. DOI: 10.1109/ICCSAI59793.2023.10421601.

[51] H. Dibeklioğlu, E. Surer, A. A. Salah, and T. Dutoit, "Behavior and usability analysis for multimodal user interfaces", *Journal on Multimodal User Interfaces*, vol. 15, pp. 335–336, 2021, Published: 16 March 2021, Issue Date: December 2021. DOI: 10.1007/s12193-021-00372-0. [Online]. Available: https://doi.org/10.1007/s12193-021-00372-0.

[52] T. Cichon and J. Rossmann, "Simulation-based user interfaces for digital twins: Pre-, in-, or post-operational analysis and exploration of virtual testbeds", in *Proceedings of the 31st European Simulation and Modelling Conference (ESM)*, Lisbon, Portugal: EUROSIS, Oct. 2017.

[53] B. R. Barricelli and D. Fogli, "Digital twins in human-computer interaction: A systematic review", *International Journal of Human–Computer Interaction*, vol. 40, no. 2, pp. 79–97, 2024. DOI: 10.1080/10447318.2022.2118189. eprint: https://doi.org/10.1080/10447318.2022.2118189. [Online]. Available: https://doi.org/10.1080/10447318.2022.2118189.

[54] M. Xia, H. Shao, X. Wang, *et al.*, "Intelligent fault diagnosis of machinery using digital twin-assisted deep transfer learning", *Reliability Engineering and System Safety*, vol. 215, p. 107 938, 2021. DOI: 10.1016/j.ress.2021.107938. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832021004531.

[55] Y. Xu, Y. Sun, X. Liu, and Y. Zheng, "A digital-twin-assisted fault diagnosis using deep transfer learning", *IEEE Access*, vol. 7, pp. 19 990–19 999, 2019. DOI: 10.1109/ACCESS.2018.2890566. [Online]. Available: https://ieeexplore.ieee.org/document/8598879.

[56] M. Segovia and J. Garcia-Alfaro, "Design, modeling and implementation of digital twins", *Sensors*, vol. 22, no. 14, 2022, ISSN: 1424-8220. DOI: 10.3390/s22145396. [Online]. Available: https://www.mdpi.com/1424-8220/22/14/5396.

[57] S. Zayed, G. Attiya, and A. El-Sayed, "An efficient fault diagnosis framework for digital twins using optimized machine learning models in smart industrial con-

trol systems", *International Journal of Computational Intelligence Systems*, vol. 16, no. 69, 2023. DOI: 10.1007/s44196-023-00241-6. [Online]. Available: https://doi.org/10.1007/s44196-023-00241-6.

[58] F. Hodavand, I. J. Ramaji, and N. Sadeghi, "Digital twin for fault detection and diagnosis of building operations: A systematic review", *Buildings*, vol. 13, no. 6, 2023. DOI: 10.3390/buildings13061426. [Online]. Available: https://www.mdpi.com/2075-5309/13/6/1426.

[59] H. Wang, J. Zheng, and J. Xiang, "Online bearing fault diagnosis using numerical simulation models and machine learning classifications", *Reliability Engineering and System Safety*, vol. 234, p. 109 142, 2023. DOI: 10.1016/j.ress.2023.109142. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832023000571.

[60] M. Islam, F. Khan, S. Alam, and M. Hasan, "Artificial intelligence in software testing: A systematic review", in *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*, 2023, pp. 524–529. DOI: 10.1109/TENCON58879.2023.10322349.

[61] Q. V. Liao, J. Wang, H. Subramonyam, and J. W. Vaughan, "Designerly understanding: Information needs for model transparency to support design ideation for ai-powered user experience", in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany: ACM, 2023. DOI: 10.1145/3544548.3580652. [Online]. Available: https://doi.org/10.1145/3544548.3580652.

[62] A. Rai, "Explainable ai: From black box to glass box", *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137–141, 2020. DOI: 10.1007/s11747-019-00710-5.

[63] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning", *arXiv preprint arXiv:1702.08608*, 2017. DOI: 10.48550/arXiv.1702.08608.

[64] S. Amershi, D. Weld, M. Vorvoreanu, *et al.*, "Guidelines for human-ai interaction", in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–13. DOI: 10.1145/3290605.3300233.

[65] M. Fan, X. Yang, T. Yu, Q. V. Liao, and J. Zhao, "Human-ai collaboration for ux evaluation: Effects of explanation and synchronization", *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW1, Apr. 2022. DOI: 10.1145/3512943. [Online]. Available: https://doi.org/10.1145/3512943.

[66] Q. V. Liao and J. W. Vaughan, "Ai transparency in the age of llms: A human-centered research roadmap", *arXiv*, vol. cs.HC, no. 2306.01941v2, 2024, Available at arXiv:2306.01941 [cs.HC]. DOI: 10.48550/arXiv.2306.01941. [Online]. Available: https://doi.org/10.48550/arXiv.2306.01941.

[67] A. B. Haque, A. N. Islam, and P. Mikalef, "Explainable artificial intelligence (xai) from a user perspective: A synthesis of prior literature and problematizing avenues for future research", *Technological Forecasting and Social Change*, vol. 186, p. 122 120, 2023, ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2022.122120. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162522006412.

[68] J. Petch, S. Di, and W. Nelson, "Opening the black box: The promise and limitations of explainable machine learning in cardiology", *Canadian Journal of Cardiology*, vol. 38, no. 2, pp. 204–213, 2022, Focus Issue: New Digital Technologies in Cardiology, ISSN: 0828-282X. DOI: https://doi.org/10.1016/j.cjca.2021.09.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0828282X21007030.

[69] K. Yu, Y. Xiao, M. Li, *et al.*, "Design for ai-integrated design team collaboration: A strategy and exploration using node flow in establishing a reusable representation of knowledge in the collaborative process", in *DRS2024: Proceedings of the DRS Biennial Conference*, C. Gray, E. Ciliotta Chehade, P. Hekkert, L. Forlano, P. Ciuccarelli, and P. Lloyd, Eds., Licensed under Creative Commons Attribution-NonCommercial 4.0 International License, Boston, USA, 2024. DOI: 10.21606/drs.2024.985.

[70] R. Moro-Visconti, S. Cruz Rambaud, and J. López Pascual, "Artificial intelligence-driven scalability and its impact on the sustainability and valuation of traditional firms", *Humanit Soc Sci Commun*, vol. 10, no. 795, 2023. DOI: 10.1057/s41599-023-02214-8. [Online]. Available: https://doi.org/10.1057/s41599-023-02214-8.

[71] Y. Xu, X. Liu, X. Cao, *et al.*, "Artificial intelligence: A powerful paradigm for scientific research", *The Innovation*, vol. 2, no. 4, p. 100 179, 2021, ISSN: 2666-6758. DOI: https://doi.org/10.1016/j.xinn.2021.100179. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666675821001041.

[72] R. Sharma, P. Agarwal, and A. Arya, "Natural language processing and big data: A strapping combination", in *New Trends and Applications in Internet of Things (IoT) and Big Data Analytics*, ser. Intelligent Systems Reference Library, R. Sharma and D. Sharma, Eds., vol. 221, Springer, Cham, 2022, ISBN: 978-3-030-99328-3. DOI: 10.1007/978-3-030-99329-0_16. [Online]. Available: https://doi.org/10.1007/978-3-030-99329-0_16.

[73] O. Alhadreti, "A comparison of synchronous and asynchronous remote usability testing methods", *International Journal of Human–Computer Interaction*, vol. 38, no. 3, pp. 289–297, 2022. DOI: 10.1080/10447318.2021.1938391. eprint: https://doi.org/10.1080/10447318.2021.1938391. [Online]. Available: https://doi.org/10.1080/10447318.2021.1938391.

[74] S. Gannouni, K. Belwafi, A. Aledaily, H. Aboalsamh, and A. Belghith, "Software usability testing using eeg-based emotion detection and deep learning", *Sensors*, vol. 23, no. 11, p. 5147, 2023. DOI: 10.3390/s23115147. [Online]. Available: https://www.mdpi.com/1424-8220/23/11/5147.

[75] E. Kuang, E. Jahangirzadeh Soure, M. Fan, J. Zhao, and K. Shinohara, "Collaboration with conversational ai assistants for ux evaluation: Questions and how to ask them (voice vs. text)", in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. DOI: 10.48550/arXiv.2303.03638. [Online]. Available: https://doi.org/10.48550/arXiv.2303.03638.

[76] M. L. van Eck, E. Markslag, N. Sidorova, A. Brosens-Kessels, and W. M. van Der Aalst, "Data-driven usability test scenario creation", in *7th International Conference on Human-Centred Software Engineering (HCSE)*, Sophia Antipolis, France, 2018, pp. 88–108. DOI: 10.1007/978-3-030-05909-5_6. [Online]. Available: https://inria.hal.science/hal-02270717.

[77] G. Bansal, T. Wu, J. Zhou, *et al.*, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance", in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966. DOI: 10.1145/3411764.3445717. [Online]. Available: https://doi.org/10.1145/3411764.3445717.

[78] N. Radziwill and M. Benton, "Evaluating quality of chatbots and intelligent conversational agents", *Data-Driven Usability Test Scenario Creation*, Apr. 2017. DOI: 10.48550/arXiv.1704.04579.