Can AI Judge Usability? A Comparative Analysis of Generative Tools on Climate Conference Websites

Danilo Teixeira Lima[1], **Alana M. Medeiros**[1], **Rita de Cássia R. Paulino**[2], **Marcos César R. Seruffo**[1]

[1]Federal University of Pará (UFPA)
Caixa Postal 479 – 66.075-110 – Belém – PA – Brasil

[2]Federal University of Santa Catarina (UFSC)
Florianópolis, SC – Brasil.

danilo.teixeira@itec.ufpa.br, alanamirandaufpa@gmail.com,

rcpauli@gmail.com, seruffo@ufpa.br

Abstract. Introduction*: This paper analyzes the usability of COP30 (Brazil) and COP29 (Azerbaijan) websites using four Generative Artificial Intelligences (GAIs): ChatGPT, DeepSeek, Gemini, and Microsoft 365 Copilot.* ***Objective:*** *To examine the role of GAIs as support tools in usability evaluations.* ***Methodology****: Following ISO 9241-210, each GAI received a standardized prompt, and outputs were analyzed through Directed Categorical Content Analysis.* ***Results:*** *GAIs identified issues such as deep navigation, absent accessibility mechanisms, and terminological inconsistencies, but showed limits in cultural sensitivity and technical accessibility. The study highlights opportunities and cautions for their application in real projects.*
Keywords *Usability, Generative Artificial Intelligence, Human-Centered Design, COP Conferences.*

1. Introduction

The digital experience is important in mediating between global audiences and international events [Castells 2008]. In particular, digital platforms such as the official websites of the Conferences of the Parties (COPs) on climate change have become not only information channels but also spaces for mobilization and political engagement. The usability of these platforms has a direct impact on users' ability to access content [Oliveira e Ferreira 2022], understand agendas, and participate in actions promoted by the events, especially considering the socio-cultural diversity and technological infrastructure of the audiences involved. In this context, evaluating the usability of global event websites has become a strategic necessity.

In this sense, the motivation for this paper comes from observing the importance of digital communication for COPs, especially given the expectation of high traffic and engagement that the COP30 website, to be held in Belém, capital of the state of Pará, in the Brazilian Amazon, should achieve. According to the Brazilian government, holding COP30 in the Amazon represents a historic opportunity for Brazil to reaffirm its leadership role in negotiations on climate change and global sustainability. It is also a strategic opportunity to give visibility to the role of Amazonian communities in the

fight against the climate crisis, highlighting their practices, knowledge, and demands at an event with international reach [Brazilian Federal Government 2024].

At the same time, this growing relevance contrasts with the persistent gaps in the literature regarding the use of accessible, replicable, and automated methodologies that can support design and communication teams in the rapid and contextualized evaluation of the usability of these environments [Kuric et al. 2025]. In light of this, this paper proposes an innovative approach by carrying out a comparative usability analysis of the COP30 and COP29 websites based on the principles of the International Organization for Standardization (ISO) 9241-210 standard, using four Generative Artificial Intelligences (GAIs): ChatGPT, DeepSeek, Gemini and Microsoft 365 Copilot. These tools, which are increasingly accessible and integrated into everyday working life, were activated through a single, standardized prompt to carry out evaluations focusing on effectiveness, efficiency, satisfaction, and suitability to the context of use.

The central question guiding the survey was: "How effectively can Generative AI tools perform comparative usability evaluations of climate conference websites, and what do their findings reveal for Human-Computer Interaction (HCI) practices?". In addressing this question, the investigation aims to: (i) to evaluate the usability of the COP30 and COP29 websites based on ISO 9241-210; (ii) to compare the results generated by the four GAIs, observing convergences, divergences, and gaps in the evaluations; and (iii) to discuss the potential for using these tools as support in HCI practice, considering their limits and possible biases.

In addition to contributing to the methodological discussion on using GAI in usability assessments, this study is also aligned with broader reflections on the implications of artificial intelligence for HCI. Specifically, it addresses the Grand Challenge GC6, as defined in the GranDIHC-BR research agenda: "Implications of Artificial Intelligence in HCI: A Discussion on Paradigms, Ethics, and Diversity, Equity and Inclusion" [Pereira et al. 2024]. By investigating how GAIs perform in evaluating digital platforms intended for global and culturally diverse audiences, this work highlights technical possibilities and critical concerns about bias, contextual insensitivity, and the risk of reinforcing exclusions in digital environments.

This study is divided into seven sections. Section 2 presents the theoretical survey that underpins the work, addressing usability on global digital platforms, traditional evaluation methods, and the emerging possibilities with the use of GAIs. Section 3 describes the methods used in the research, detailing the selection of the sites investigated, the GAI tools applied, and the procedures adopted for collecting, organizing, and analyzing the data. Section 4 presents the results obtained, organized on three fronts: analysis by usability criteria, comparison between the COP30 and COP29 sites, and discussion of the potential and limitations of GAIs. Section 5 discusses the findings of the study, relating them to the research objectives and previous work in the field. Section 6 discusses the ethical precautions adopted in the study. Finally, section 7 presents the final considerations, including the limitations of the work, threats to validity, and possibilities for future research.

2. Theoretical Survey

This section brings together the main conceptual foundations on which this study is based. It addresses the relevance of usability on digital platforms with a global reach, discusses traditional and automated evaluation methods, emphasizes the emerging use of GAIs, and presents gaps identified in the literature.

2.1. Usability on Global Digital Platforms

The digital experience on online platforms plays an increasingly central role in how global audiences interact and engage with international events of great political, environmental, and media significance, such as the COPs of the United Nations Framework Convention on Climate Change (UNFCCC). The official websites of these events go beyond the role of mere information repositories, acting as access portals to details on registration, discussion agendas, and mobilization opportunities. In a globalized context, the usability of these digital platforms becomes a determining factor in ensuring the effective participation of a diverse public, breaking down geographical barriers, and facilitating access to information and civic engagement on issues of global reach [Nielsen 1999]. High usability fosters inclusivity, enabling stakeholders from varied cultural, linguistic, and technological backgrounds to engage meaningfully with complex global issues like climate change.

The ISO 9241-210 standard, which defines usability as the extent to which a system can be used by specific users to achieve their goals effectively, efficiently, and satisfactorily within a particular context of use, provides a robust framework for evaluating and designing user-centered digital interfaces [Bevan et al. 2015]. On digital platforms such as those of the UNFCCC events, high usability translates into greater accessibility to information, simplified registration processes, and intuitive navigation, elements that are central to encouraging the participation of individuals and organizations from different parts of the world [Norman Donald 2013]. This standard emphasizes iterative design processes that incorporate user feedback, which is particularly relevant for global platforms aiming to serve heterogeneous audiences [International Organization for Standardization 2010]. The choice of ISO 9241-210 for evaluating COP websites is justified by its international recognition and its comprehensive guidelines for human-centered design, which prioritize accessibility, inclusivity, and user satisfaction in complex interactive systems.

Beyond ISO 9241-210, the research underscores the critical role of usability in global digital platforms. For instance, [Nielsen 1999] argue that effective usability design enhances user trust and engagement, which is necessary for platforms hosting high-stakes global discussions. Their work highlights how poor usability, such as confusing navigation or inaccessible content, can alienate users, particularly those from underrepresented regions or with limited digital proficiency. In the context of UNFCCC platforms, this could hinder participation from developing nations, where access to high-speed internet or advanced devices may be limited. A study by [Cyr et al. 2006] further explores the cultural dimensions of usability, demonstrating that cultural preferences, such as color schemes, layout styles, and language localization, influence user satisfaction and engagement on global websites. For COP websites, incorporating culturally sensitive design elements, such as multilingual interfaces or region-specific content formats, can enhance accessibility for diverse global audiences.

To further enhance the usability of COP websites, integrating principles from cross-cultural design and accessibility standards is essential. Research by [Shneiderman e Plaisant 2010] emphasizes the importance of designing interfaces that accommodate diverse user needs through clear visual hierarchies and adaptive content presentation, which can mitigate barriers for users with varying levels of digital literacy or technological access. Additionally, a study by [Hassenzahl e Tractinsky 2006] highlights the role of user experience in fostering emotional engagement, suggesting that aesthetically pleasing and intuitive designs can strengthen user trust and motivation to participate in global initiatives like the COPs. By prioritizing culturally responsive, accessible, and user-centered design, these platforms can effectively bridge global divides, ensuring that critical climate change discussions are inclusive and impactful for all stakeholders.

2.2. Traditional Usability Evaluation Methods

Traditional usability evaluation methods, although widely validated, can present challenges when applied to large-scale digital platforms such as global event websites. Heuristic analysis, proposed by [Nielsen 1994], is based on general principles that may not adequately capture the cultural and linguistic nuances of diverse audiences. Studies indicate that cultural and linguistic factors influence the results of usability tests, highlighting the need for approaches that are more sensitive to these variables [Vatrapu e Pérez-Quiñones 2004].

Testing with real users becomes logistically complex and costly in global contexts. As noted by [Liu 2021], international usability testing requires recruiting participants in multiple countries, taking into account cultural, technological, and connectivity variations, a process that can take months and require significant investment. The System Usability Scale (SUS), developed by [Brooke 1996], although practical, shows limitations in assessing critical aspects of complex platforms. Research shows that the SUS is less sensitive to detecting problems specific to multilingual interfaces, such as inconsistencies in the location of content or differences in the spatial organization of elements between language versions [Miraz et al. 2017]. Emerging alternatives, such as automated remote assessments and interaction data mining techniques, have been proposed to overcome these limitations. However, these approaches still need large-scale validation for contexts as complex as COPs' websites.

2.3. Related Work: The Possibilities of Generative AI for Evaluating Interfaces

Automated usability evaluation of interfaces has been the subject of study since the early 2000s, with pioneering efforts aimed at systematizing and reducing costs in the usability verification process. [Ivory e Hearst 2001] carried out extensive research into usability evaluation methods, organizing them into a new taxonomy that emphasizes the role of automation. However, traditional methods based on coded heuristics face limitations in highly complex scenarios, such as multilingual interfaces, global platforms, and applications with multiple user profiles. More recently, GAIs has opened new possibilities for automating usability assessments in more contextual and flexible ways.

Recent studies have explored the use of GAIs to support interface design tasks, generate contextualized feedback, and perform automated inspections based on consolidated usability principles. [Fischer e Lanquillon 2024] evaluated the impact of

GAI on software design and engineering through a user-centered approach, identifying critical tasks within the software development lifecycle where GAI tools can improve productivity and creativity. In addition, [Bleichner e Hermansson 2023] investigated the usefulness of GAI in the design of user interfaces, using a stable diffusion model adjusted to create images of login screens based on text prompts provided by users. The concept was evaluated through user testing, demonstrating how GAI models can translate user-provided textual prompts into functional interface components.

However, the literature still needs more robust studies that systematically validate the effectiveness of these tools when compared to established usability evaluation methods. The study by [Araújo et al. 2024] is one of the few to combine GAIs with human inspection to evaluate the usability of an institutional website, in which the Greenpeace Brazil[1] website was evaluated, using the ISO 9241-210 standard as a basis. This research highlights the complementarity between the automatic analyses generated by GAIs and the expertise of human experts, providing evidence that hybrid approaches can increase the reliability of the evaluation. Another relevant contribution is the experience report by [Borges e Araújo 2024], which explores the integration of AI tools, such as GPT-4 and Gemini 1.5, in the redesign process of the educational platform Classroom eXperience, showing how these tools can complement designers by offering precise analyses and usability suggestions.

Beyond studies that integrate GAIs into interface design and evaluation, additional research has contributed to understanding usability challenges in institutional platforms and the role of AI in user interaction. An important question in the field of interface evaluation, particularly in public and institutional contexts, is the study by [Serra et al. 2015], which conducted an accessibility evaluation of e-government mobile applications in Brazil. The authors adopted a systematic method to assess how well these apps complied with accessibility guidelines, revealing usability barriers, especially for users with disabilities. Although this study does not focus on GAIs, it demonstrates the importance of inclusive design in governmental platforms, a concern that also underpins this comparative analysis of COP29 and COP30 in this study. The findings reinforce the need for tools and approaches that can efficiently identify accessibility issues, especially in large-scale digital environments.

In a more recent context, [Alsadi e Miller 2023] investigated the impact of large language models, particularly ChatGPT, on user experience. Their study explored how AI-generated responses affect users' perceptions of relevance, engagement, and trust. The results emphasize both the potential and the limitations of GAI models in shaping digital interactions, particularly when user expectations and context-sensitive outputs are involved. Complementarily, [Li 2024] proposed optimizations for the user experience design of a campus library information system, highlighting that successful digital experiences rely not only on technical implementation but also on users' cognitive models and behaviors, a concern reflected in the ISO 9241-210 framework adopted in the study. These contributions strengthen the understanding that evaluating usability requires adaptive and culturally aware methodologies, especially when platforms serve diverse, multilingual, and global audiences.

[1]https://www.greenpeace.org/brasil/

Thus, there is still a lack of comparative studies in large-scale contexts, particularly those that consider the cultural and linguistic diversity of users. In this sense, the present study advances the literature by proposing an approach that integrates GAIs into the usability evaluation of digital interfaces. Unlike previous work, such as [Araújo et al. 2024] and [Borges e Araújo 2024], which combined GAIs with human inspection in specific institutional contexts, this investigation applies multiple GAI models in a comparative analysis of the COP30 and COP29 websites, grounded in the ISO 9241-210 usability principles. This strategy enables a broader understanding of how different design decisions impact usability across diverse user groups.

Furthermore, by addressing challenges identified in prior research, such as the technical limitations of automated heuristics [Ivory e Hearst 2001], the contextual sensitivity required in AI-supported evaluations [Alsadi e Miller 2023], and the barriers to accessibility in government platforms [Serra et al. 2015]. This study contributes to the refinement of evaluation methods that are scalable, adaptive, and attentive to inclusion. It also responds to recent demands for analytical approaches capable of operating across multilingual and heterogeneous environments [Li 2024]. In doing so, this investigation broadens the scope of usability research applied to global digital platforms, with an emphasis on methodological rigor and cultural awareness.

3. Research Methods

To conduct this study, a multi-step methodological procedure was adopted to ensure the systematic collection, organization, and analysis of the data generated by the selected GAI models. Figure 1 illustrates the general workflow and summarizes the main stages of the research process.

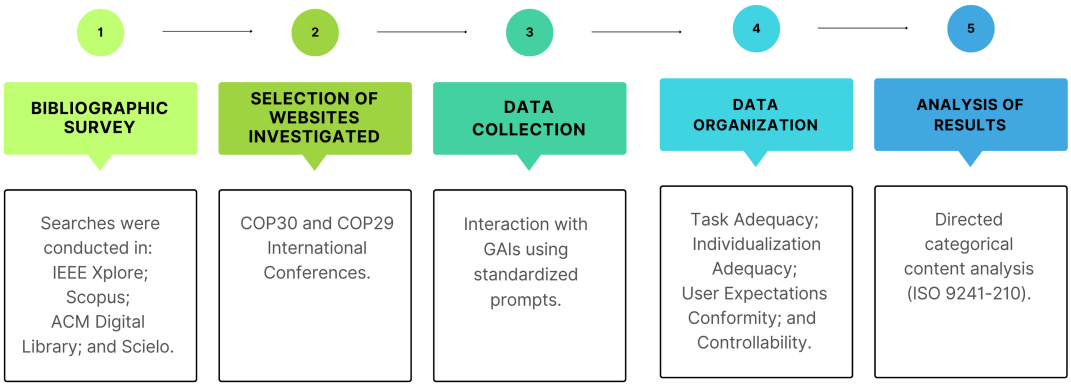

Figure 1. Methodological Path.

3.1. Bibliographic Survey

A structured bibliographic search was conducted to support the theoretical background of this study and ensure the relevance and quality of the selected works. The process included:

I) Definition of Research Focus: The central question guiding the research was: "How effectively can Generative AI tools perform comparative usability evaluations of climate conference websites, and what do their findings reveal for HCI practices?".

II) Inclusion and Exclusion Criteria: Included were studies focused on GAI applications in usability, human-computer interaction, or automated evaluation methods; excluded were works not directly addressing evaluation methods or lacking full-text availability.

III) Database Search: Searches were conducted in IEEE Xplore ([2]), Scopus ([3]), ACM Digital Library ([4]), and Scielo ([5]), using keywords such as Generative AI, Usability Evaluation, Automated Heuristic Evaluation, and Human-Computer Interaction.

IV) Initial Screening: Studies were initially selected based on titles and abstracts, assessing their relevance to the research focus.

V) Full-Text Review and Final Selection: The selected works underwent full-text review, evaluating methodological alignment with the study's goals.

VI) Qualitative Analysis and Synthesis: Relevant findings were extracted and organized to inform the methodological framework and discussion of results.

3.2. Selection of Webites Investigated

For the usability analysis, the websites of the COP30[6] and COP29[7] international conferences were selected, both in their English versions, as events of great global relevance in the discussions on climate change. The COP30 website was chosen because it was historically held in Brazil, which adds a symbolic and strategic layer to the event, with a particular focus on the Amazon and the region's environmental issues. The COP29 website was included to allow a comparison between the two platforms, highlighting geopolitical and contextual differences that can impact the user experience.

Although COP29 also dealt with global climate issues, its holding in another context, in Azerbaijan, situated at the crossroads between Eastern Europe and Southwest Asia, offers a unique opportunity to contrast with the COP30 website, facilitating the analysis of more effective design strategies for a diverse global audience. The comparison aims to identify good practices and suggest improvements, especially about inclusion and accessibility on large-scale platforms. Both websites were analyzed in their desktop versions between March 10 and 15, 2025, with a focus on navigation structure, access to official documents, and public engagement functionalities.

To visually contextualize the objects of study, Figure 2 shows the homepage of the COP30 website, while Figure 3 the homepage of the COP29. The images are intended to provide an overview of the interfaces analyzed, helping to understand the visual and structural scope of each platform.

3.3. GAIs Used and Standard Prompt

For the usability analysis of the COP30 and COP29 websites, four GAI models were used, all with the "think deeper" option activated to ensure a more detailed analysis.

[2]https://ieeexplore.ieee.org/xplore/home.jsp
[3]https://www.scopus.com/
[4]https://dl.acm.org/
[5]https://scielo.br/
[6]https://cop30.br/en/
[7]https://cop29.az/en/

Figure 2. Homepage of the COP30 Website.

The GAIs used were: ChatGPT (OpenAI)[8] - version 4 turbo; DeepSeek[9] - version 3; Gemini[10] (Google) - version 2.0 flash; and Microsoft 365 Copilot[11] - latest version (no specific version number). All the models were used in the English version, and a single standardized prompt was applied to all the GAIs in order to guarantee consistency in the analyses. The standardized prompt used was:

> "Make a detailed usability analysis of the website COP30 (URL: `https://cop30.br/en`) based on the requirements of the ISO 9241-210 standard, which addresses effectiveness, efficiency, and user satisfaction. The analysis should include specific evaluations on:
>
> - Effectiveness: ability to fulfill objectives;
> - Efficiency: time and effort required to accomplish tasks;
> - Satisfaction: subjective perception of the user;
> - Related heuristics.
>
> Repeat the same analysis for the website COP29 (URL: `https://cop29.az/en`).
>
> Then, directly compare the usability of the two sites, highlighting:
>
> - Strengths and weaknesses of each;
> - Significant similarities and differences;
> - Concrete recommendations for improvements to the COP30 site, considering the context of a global audience engaged in climate change.
>
> Format the answer in three clear sections: Analysis of COP30, Analysis of COP29, and Comparison and Recommendations."

[8]https://chatgpt.com/
[9]https://chat.deepseek.com/
[10]https://gemini.google.com/
[11]https://copilot.microsoft.com/

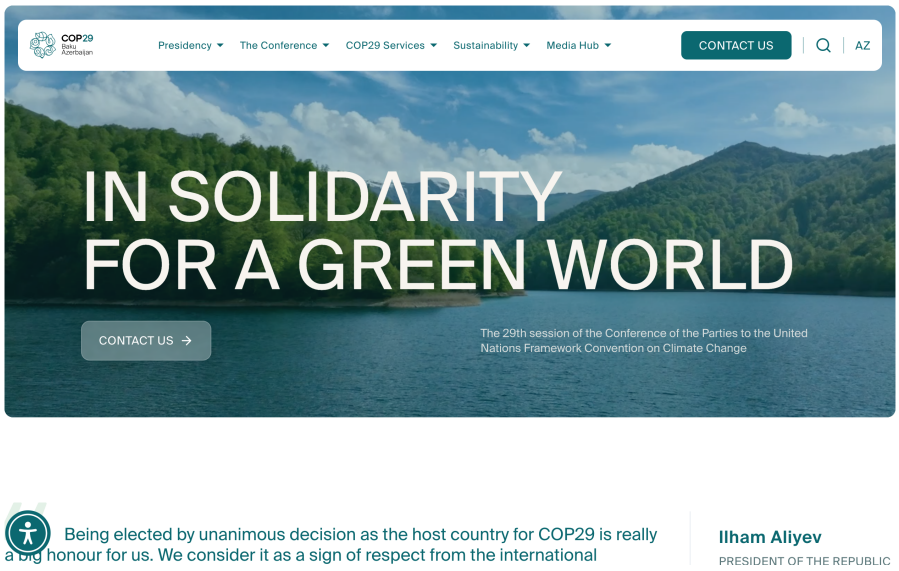

Figure 3. Homepage of the COP29 Website.

This command directed the GAIs to provide a more detailed and standardized analysis, focusing on the three main issues of ISO 9241-210: Effectiveness (delivering conference information and engaging users); Efficiency (time/effort required for navigation and information retrieval); and Satisfaction (users' subjective experience regarding ease of use, content clarity, and overall platform experience) [Bevan et al. 2015]. In addition, the GAIs were instructed to directly compare the COP30 and COP29 websites, highlighting the strengths and weaknesses of each, their similarities and significant differences, and offering recommendations for improving the COP30 website.

3.4. Procedures

The collection, organization, and comparison of the responses generated by the GAIs followed a protocol aligned with the principles of ISO 9241-210 and the requirements of a structured analysis:

I) Data Collection: The analysis was carried out through interactions with each of the selected GAIs. Standardized prompts were inserted into the platforms of each tool, which generated responses on the usability of each website, according to the criteria established by ISO 9241-210. Interactions with the GAIs were documented, ensuring that the collection process was replicable and consistent between the different tools. Each GAI was used under the same conditions and with the same prompt to ensure the comparability of responses.

II) Data Organization: Followed a categorization method based on the main criteria of ISO 9241-210, the criteria being Task Adequacy (Evaluation of the website's efficiency in enabling the user to achieve their objectives); Individualization Adequacy (Verification of personalization and adaptation to the user's needs); User Expectations Conformity (Analysis of consistency with expected navigation patterns and terminology); and Controllability (Degree of control the user has over the interaction. All the responses generated by the GAIs were grouped according to these criteria, allowing for a structured

evaluation of the platforms).

III) Analysis of Results: After the data organization phase, a Directed Categorical Content Analysis was conducted, following the principles proposed by [Bardin 2011]. This technique, grounded in qualitative research, involves predefined categories derived from theoretical frameworks, in this case, the ISO 9241-210 usability criteria. Unlike purely inductive approaches, the directed strategy allows the analysis to be guided by prior conceptual structures, which is particularly advantageous when the goal is to verify how well existing theoretical constructs manifest in new empirical contexts.

The choice of this method is justified by the need to systematically interpret unstructured qualitative data generated by different GAIs tools, ensuring consistency across sources. It also enabled a comparative examination of how each GAI addressed the same usability dimensions. The categorization process was performed manually through analytical reading of the full response reports and the coding of relevant excerpts into a structured spreadsheet, allowing both frequency analysis and interpretive insights.

To ensure analytical precision, a selective coding strategy was adopted to avoid redundancies commonly found in generative model outputs, where semantically equivalent phrases may be repeated in various forms. Rather than counting every mention aligned with a usability criterion, only salient, contextually distinct excerpts were considered. This approach reduced the risk of inflated frequencies and better reflected the diversity of the observations provided by each GAI, following the principles of targeted content analysis as outlined by [Bardin 2011].

4. Results

The results section presents the main findings obtained from the analysis of the responses generated by the GAIs, organized on the basis of the ISO 9241-210 criteria. The findings are discussed on three fronts: the analysis by usability criterion; the comparison between the COP30 and COP29 websites; and the evaluation of the potential and limitations of GAIs as tools applied to usability analysis. Each excerpt quoted in this section is accompanied by an identifier (ID), which makes it possible to trace its origin in the qualitative database. The complete spreadsheet with the systematized categorization of all the data collected is available on the Zenodo platform[12].

4.1. Analysis by Criteria

Given the qualitative nature of the data, coming from extensive responses generated by four GAIs (ChatGPT, DeepSeek, Gemini, and Microsoft 365 Copilot), we opted to apply a Directed Categorical Content Analysis [Bardin 2011]. The answers were extracted and organized into three blocks: i) Evaluation of the COP30 website by each GAI; ii) Evaluation of the COP29 website by each GAI; and iii) Comparison and recommendations between the two websites. Each relevant excerpt was classified according to ISO criteria and categorized manually, allowing for a quantitative (frequency of mentions) and qualitative (interpretation of patterns) analysis.

To avoid redundancy and overrepresentation of repeated concepts, a common phenomenon in GAI results, a selective coding strategy was adopted based on the

[12]https://zenodo.org/records/15307148

principles of targeted content analysis, in which only salient and contextually distinct passages were counted for comparative frequency [Bardin 2011]. This approach made it possible to avoid distortions that would arise from counting paraphrased or semantically equivalent mentions in the same GAI result. Therefore, the frequencies reported in this section do not represent raw totals of all phrases aligned with each ISO criterion but rather the number of unique and analytically relevant mentions per model, criterion, and website.

Figure 4 provides a comparative overview of the frequency of usability findings reported by each GAI across the evaluated criteria, based on their analysis of the COP30 and COP29 websites. The visualization highlights consistent patterns as well as notable differences in how each model interprets and applies the usability principles defined by ISO 9241-210.It is important to note that the frequency values summarized are not derived from the raw total of all categorizations present in the complete spreadsheet, but were constructed based on the synthesis of salient, non-redundant findings identified across the GAI responses. In contrast, the full table aggregates all coded entries, including overlapping or semantically similar feedback, and therefore serves a different descriptive purpose.

Figure 4. Graph of Frequency of Mentions by Usability Criterion.

About the **Task Adequacy** criterion, COP29 was rated most positively by all the GAIs, with 5 mentions each, while COP30 received between 3 and 5 mentions, depending on the tool. According to the GAIs' outputs, COP29 was described as having a more straightforward structure, with quicker access to the program and logistical information. For example, ChatGPT reported, *"The homepage directs users to key areas such as registration, schedule, and logistics."* (ID: 196), while DeepSeek observed for COP30 that *"The main navigation is generally clear but certain submenus require multiple clicks, reducing task efficiency."* (ID: 23). These outputs suggest that COP29 was perceived by the GAIs as more efficient in enabling task completion, whereas COP30 demanded more cognitive effort due to less intuitive navigation.

Regarding the criterion **Individualization Adequacy**, this was the least frequently mentioned aspect across all evaluations, with only 1 or 2 mentions per GAI on both websites. This low frequency reflects both the limited presence of explicit personalization features and the analytical decision to count only conceptually distinct passages from each tool's output. For instance, repetitive mentions of the same missing accessibility or customization feature were consolidated to avoid inflating the data. The GAI analyses indicated a general lack of personalization on both sites, including the absence of interface

customization and user preference options. Copilot, for example, noted, *"No available options to change font sizes or activate dyslexic-friendly fonts"* (ID: 138), and Gemini reported, *"No user-specific dashboard or profile customization options are offered"* (ID: 88). These evaluations point to a shared limitation in both sites, revealing a gap in user-centered design, as defined by ISO 9241-210, which advocates for adaptable and flexible systems.

When analyzing the criterion **User Expectations Conformity**, the evaluations produced by the GAIs once again favored COP29, which received between 3 and 4 mentions per tool, while COP30 received 2 to 3. The GAIs indicated that COP29 follows more intuitive standards, such as clear terminology and expected organization. For example, Gemini stated, *"Navigation menus are logical but submenus sometimes use vague labeling that could confuse users"* (ID: 87). In contrast, COP30 was frequently criticized by the GAIs for terminological inconsistencies and dispersion of related content, *"Terminology inconsistency, such as the coexistence of 'Programme' and 'Agenda', may confuse international visitors"* (ID: 6). These findings indicate that COP29 was more aligned with user expectations, whereas COP30 presented barriers that may have hindered the overall user experience.

In the last criterion, **Controllability**, both websites received criticism for the lack of mechanisms to facilitate user control, such as breadcrumbs or 'back to top' buttons. COP29, however, had slightly better ratings (2-3 mentions) compared to COP30 (1-2 mentions). DeepSeek, for instance, noted, *"The site does not provide a 'Back to Top' button on longer pages, requiring manual scrolling."* (ID: 26) and Copilot stated, *"Long pages lack sticky headers or quick navigation tools, complicating content traversal."* (ID: 139). These evaluations reinforce the interpretation that controllability was not a strong focus in either wesite's design.

4.2. Comparison of Websites: COP30 vs COP29

A comparison between the COP30 and COP29 websites reveals both basic convergences and more divergences in their design approaches. Both share a similar fundamental structure, organizing essential information such as logistics (visas, accommodation, and transport) in dedicated sections and using thematic navigation as a central organizing principle. This structural similarity suggests the adoption of established conventions for large-scale event websites. As part of the evaluation process, it is worth noting that the response time of the GAIs when processing the same prompt varied significantly: ChatGPT took 34 seconds, DeepSeek 36 seconds, Copilot 35 seconds, and Gemini required 4 minutes and 21 seconds. While not directly related to the content of the analysis, this variation reflects differences in system efficiency that may become relevant in iterative usability evaluation workflows.

Following the structural overview, a qualitative analysis was conducted based on the responses generated by the GAIs. Table 1 presents a synthesis of representative usability findings extracted from these evaluations. The table contrasts key aspects of the COP30 and COP29 websites across five usability dimensions: Navigation and structure; Visual design; Accessibility, Terminology and expectations; and User control. Each observation is supported by a traceable excerpt from the GAIs' outputs (indicated by their respective IDs). This comparative summary introduces the main differences identified between the two platforms and supports the detailed analysis that follows.

Table 1. Synthesis of usability highlights for COP30 and COP29 based on GAI evaluations.

Evaluation Dimension	COP30	COP29
Navigation and structure	"The site structure aligns well with the expectation of finding logistics information." (ChatGPT, ID: 9); "Submenus require multiple clicks, reducing task efficiency." (DeepSeek, ID: 23)	"The homepage clearly directs users to key areas such as registration, schedule, and logistics." (ChatGPT, ID: 196); Shallower menus and clearer top-level items
Visual design	"Homepage banners are visually appealing but may distract from the main calls to action." (DeepSeek, ID: 28); "Call-to-action buttons blend into background images." (Copilot, ID: 156)	No specific highlights for aesthetic impact; evaluations focused on clarity and structure
Accessibility	"No accessibility widget visible." (ChatGPT, ID: 12); "Contrast ratios in some banner images do not meet accessibility standards." (Copilot, ID: 153)	"Accessibility options are not clearly advertised." (ChatGPT, ID: 208); Minor accessibility concerns noted
Terminology and expectations	"Terminology inconsistency, such as the coexistence of 'Programme' and 'Agenda'." (ChatGPT, ID: 6)	"Logical menu structure with predictable grouping." (Gemini, ID: 95); "Aligns well with user mental models." (Copilot, ID: 195)
User control	"No sticky headers or 'Back to Top' buttons." (ChatGPT, ID: 10); "Carousel content cannot be paused." (Gemini, ID: 89)	"Slightly better control tools, but no advanced features." (Copilot, ID: 198)

Differences in the implementation of these basic principles result in different experiences.According to ChatGPT's output, *"The site structure aligns well with the expectation of finding logistics information under a clear 'Logistics' section"* (ID: 9) for COP30, while for COP29 it reported, *"The homepage clearly directs users to key areas such as registration, schedule, and logistics."* (ID: 196). Based on this, COP29 was described as having greater efficiency in the navigation architecture, with shallower menus and more obvious main items. This approach, which follows consolidated recommendations in HCI literature on depth versus breadth in navigation, was interpreted by the GAI as facilitating quicker access to critical information.

On the other hand,the GAIs evaluations suggested that COP30 invested more heavily in visual identity elements, creating an aesthetically richer experience. However, this emphasis on the visual aspect was not accompanied by corresponding improvements in practical usability. COP30 invested more heavily in visual identity elements, creating an aesthetically richer experience. However, this emphasis on the visual aspect has not translated into corresponding improvements in practical usability. On the contrary, analysis indicates that the quest for visual differentiation may have compromised

important aspects of intuitive navigation. According to the GAI analysis, the effort toward visual differentiation may have compromised important aspects of intuitive navigation.

For instance, Copilot noted that "call-to-action buttons blend visually into background images, reducing their visibility" (ID: 156), and "contrast ratios in some banner images do not meet accessibility standards for readability" (ID: 153), indicating that visual refinements compromised functional clarity. Similarly, DeepSeek observed that "the homepage banners are visually appealing but may distract from the main calls to action" (ID: 28), and ChatGPT reported that "the visual hierarchy generally supports task completion, but occasional inconsistent font sizing can distract users" (ID: 18). These examples suggest that although COP30 presented a more polished visual identity, the visual enhancements were not accompanied by improvements in usability and, in some cases, hindered effective interaction.

A particularly relevant contrast concerns accessibility practices. According to the ChatGPT output, *"The absence of an accessibility widget visible (e.g., text-to-speech, color adjustment tools)."* (ID: 12) was noted for COP30, while for COP29 it stated, *"Accessibility options such as screen reader optimization or high contrast modes are not clearly advertised"* (ID: 208). These observations reflect different levels of alignment with inclusive design principles, which are critical for events aimed at global and diverse audiences.The comparison of GAI responses suggests that, while both websites share a common base of functionality, the implementation choices resulted in different qualitative experiences.

The GAIs evaluated COP29 more favorably in terms of operational efficiency and inclusiveness, whereas COP30 emphasized branding and visual identity elements. This divergence reveals how distinct design priorities can shape user experience in different ways. While COP29 appears to favor clarity, accessibility, and direct access to key content, COP30 invests na construção de uma imagem institucional mais marcante, ainda que com eventuais compromissos na usabilidade. The synthesis presented in Table 1 illustrates how these strategic choices are reflected across core usability dimensions, offering a comparative snapshot of the platforms' strengths and limitations.

4.3. Potential and Limitations of GAIs in Usability Evaluation

The results show that GAIs have promising potential as complementary tools in usability evaluation processes, although they also reveal notable limitations. The comparative analysis of the COP30 and COP29 websites, conducted solely through GAI evaluations, revealed that three of the four GAIs evaluated (ChatGPT, DeepSeek, and Copilot) were consistent in identifying usability problems.

These tools converged in their observations on problems related to navigation structure, although with varying emphases. Among the most relevant findings was the identification of the excessively deep navigation structure in COP30 in contrast to the flatter navigation of COP29. For example, ChatGPT reported, *"Navigation menus are intuitive and group related information effectively, although deeper pages require several clicks to access"* (ID: 2); DeepSeek observed, *"The main navigation is generally clear but certain submenus require multiple clicks, reducing task efficiency"* (ID: 23); and Copilot noted, *"Long pages lack sticky headers or quick navigation tools, complicating content traversal"* (ID: 139). These GAIs also acknowledged the absence of a section

dedicated to accessibility at COP30, noted by three of the four tools.

These evaluations suggest that GAIs, when prompted with structured heuristics, can effectively detect patterns in information architecture, positioning them as valuable tools for preliminary usability assessments, especially in large-scale digital projects. One illustrative example was the identification by ChatGPT and Copilot of content fragmentation in COP30's media section, where videos were categorized under "News" and images under "Press", a practice inconsistent with the principle of content grouping and visual consistency. Such capacity for systematic scanning and classification could be especially helpful in the early triage phases of iterative redesigns, allowing professionals to identify and prioritize critical issues before human testing is conducted.

However, the case of Gemini highlighted an important constraint in GAI based evaluation: their strong dependency on accessible input. Unlike the other tools, Gemini returned a "currently inaccessible" message when prompted to analyze the site, failing to generate insights due to an inability to access or adapt to the page. This episode demonstrates a broader limitation of current GAIs, they are not autonomous systems and are unable to devise workaround strategies when facing external technical barriers. This limits their reliability in remote usability evaluations and may introduce false negatives when content is temporarily or regionally unavailable.

Further limitations also emerged regarding the technical and subjective dimensions of the evaluation. None of the GAIs produced thorough assessments of core accessibility features, such as contrast ratio, screen reader compatibility, or keyboard navigation, despite their central role in ISO 9241-210. These areas still require manual verification and expert analysis. A tendency toward superficiality was also observed: the tools focused on more conspicuous usability problems (e.g., menu depth), often overlooking contextual or cultural dimensions of user experience. Additionally, inconsistency became evident in evaluations related to user satisfaction. Different GAIs offered contradictory perspectives on identical design elements, underscoring the current limitations in simulating nuanced human perception and judgment.

5. Discussions

The findings of this study reveal an ambiguous yet promising scenario regarding the application of GAIs in the usability evaluation of digital platforms. On the one hand, the data indicates that these tools are capable of carrying out structured and coherent diagnoses, offering relevant pointers for the initial stages of evaluation. Nonetheless, there are considerable limitations relating to the depth of analysis, cultural sensitivity, and adaptability to real contexts of use.

The four GAIs analyzed demonstrated, with variations, a remarkable ability to identify recurring usability patterns. Problems such as deep navigation structures, the absence of control elements (such as breadcrumbs or back-to-top buttons), flaws in the organization of content, and terminological inconsistency were recurrent in the analysis of the tools. This suggests that, when subjected to well-formulated and consistent prompts, GAIs can operate as "automated heuristic auditors", capable of pointing out the most obvious and recurring weaknesses in the websites evaluated.

This finding resonates with early research by [Ivory e Hearst 2001], who classified automated usability evaluation methods and emphasized their potential to reduce costs

and increase efficiency. However, that same literature already warned of limitations in complex contexts, such as multilingual systems and diverse user profiles, challenges that remain evident in the GAI-based evaluations conducted in this study.

More recently, [Bleichner e Hermansson 2023] explored how GAI can be used to support interface creation through text-based prompts, reinforcing its value in early design stages. While those studies focus on generation and prototyping, this paper expands the discussion to the evaluation phase, testing GAIs not as design assistants, but as autonomous usability reviewers operating on real websites under consistent conditions. A relevant example is provided by [Borges e Araújo 2024], who examined the experiences and challenges of integrating an AI assistant into the redesign process of an educational platform, emphasizing the complementary role of GAI in generating precise usability suggestions.

This competence of GAIs aligns not only with earlier research into automation but also with current trends in HCI. For instance, [Fischer e Lanquillon 2024] highlights the potential of these tools to automate analytical tasks and support productivity in design cycles. The findings of this study also corroborate the work of [Araújo et al. 2024], who demonstrated that GAIs can effectively support human inspection, particularly in preliminary usability diagnostics. Building upon this perspective, the present research advances by testing the autonomy of GAIs as primary evaluators, enabling the identification of comparative response patterns across different tools, something that has not been explored in the literature.

These results also demonstrate broader concerns identified in the literature regarding the use of digital platforms by heterogeneous audiences. [Serra et al. 2015], for example, had already warned about persistent barriers in public applications, especially in relation to accessibility, a point partially captured by GAIs in this study, but not always addressed with sufficient depth or contextual nuance. Likewise, while [Alsadi e Miller 2023] points to the potential of language models to improve interaction, their findings also highlight user concerns with consistency and clarity, which reinforces the need for prompt calibration and post-hoc verification when applying GAIs to usability tasks. From another perspective, [Li 2024] draws attention to the importance of aligning interface design with users' mental models, something GAIs struggle to assess effectively, given their limited understanding of user diversity, the context of access, or local design expectations. These limitations suggest that, despite their efficiency, GAIs should be applied with caution in institutional evaluations, especially when inclusive design and cross-cultural dimensions are critical.

However, the limits also became evident. The Gemini GAI, for example, failed to access the COP30 website, returning a generic response of inaccessibility ("currently inaccessible"), which highlights a relevant technical limitation, the tool's inability to adapt to errors or create alternative strategies to obtain the information. This behavior can compromise large-scale remote evaluations, where access can be unstable or partial, and shows that the reliability of the analysis depends not only on the tool but also on the technical context of the environment being evaluated.

Another point of attention identified in this study refers to the limited depth with which GAIs approached subjective dimensions such as user satisfaction, aesthetic

clarity, and emotional perception of the interface. In many instances, the tools offered contradictory evaluations of the same interface elements, suggesting that, unlike structural or functional dimensions, where greater alignment was observed, subjective analysis still presents significant inconsistencies across different GAIs. This pattern of divergence is illustrated through various excerpts extracted from the GAI evaluations and exemplifies not only different emphases in judgment but also the difficulty in establishing a consistent subjective interpretation across GAI outputs.

To illustrate these inconsistencies more clearly, Table 2 summarizes how each GAI assessed selected interface elements of the COP30 website. The side-by-side comparison reveals both overlapping insights and interpretive divergences, reinforcing the observation that subjective dimensions of usability, especially those linked to visual experience, are not yet reliably captured by generative models. This suggests that caution should be exercised when employing GAIs to assess non-quantifiable aspects of digital design.

Table 2. Comparison of GAI Evaluations of Selected COP30 Website Elements

Element evaluated (COP30)	**ChatGPT**	**DeepSeek**	**Gemini**	**Copilot**
Visual hierarchy	"Supports task completion, but inconsistent font sizing." (ID: 18)	"Homepage banners may distract from calls to action." (ID: 28)	"Relies on large banners, pushing content below fold." (ID: 98)	—
Navigation structure	"Navigation menus are intuitive." (ID: 2)	"Submenus require multiple clicks." (ID: 23)	"Menus are logical but vague labeling." (ID: 87)	"Good first-level navigation, deeper layers could improve." (ID: 137)
Content placement (homepage)	"Homepage offers clear overview." (ID: 1)	"Key event dates are in nested pages." (ID: 25)	"Homepage highlights key info." (ID: 86)	"Essential content accessible, but urgent items lack emphasis." (ID: 136)
Call-to-action visibility	"Visual hierarchy inconsistent." (ID: 18)	—	"Buttons blend into background." (ID: 121)	"CTAs blend visually, reducing visibility." (ID: 156)
Accessibility visual tools	"No accessibility widget visible." (ID: 12)	"Missing alternative text." (ID: 41)	"Accessibility statement lacks detail." (ID: 101)	"Contrast ratios do not meet standards." (ID: 153)

The comparative layout further reveals that, while there is general alignment among GAIs in identifying structural usability issues, such as primary navigation flows, substantial divergence remains in the analysis of subjective and interpretive elements, including content prominence and visual balance. Although all tools recognized core strengths and weaknesses, their emphases, descriptive language, and analytical angles differed. This variation reinforces the argument that GAIs, while useful in supporting usability evaluations, must be triangulated and critically reviewed, especially when applied to human-centered design. Beyond technical insights, these findings help refine

methodological practices for incorporating generative models into structured evaluation workflows.

Finally, this study contributes to the scientific community by proposing and testing a standardized usability evaluation model based on GAIs, using the ISO 9241-210 criteria and applying them in a comparative way between different tools. The research offers a systematic analysis of the potential and limitations of these technologies, demonstrating their application in a real, complex, and sensitive context, the websites of the UN COPs, marked by high international visibility, political relevance, and demands for digital equity. In addition, the study advances the methodological discussion by exploring the systematized use of multiple GAIs as complementary instruments in the comparative evaluation of usability.

6. Ethical Care

This research did not involve the direct participation of human beings, nor the collection of personal or sensitive data. For this reason, it was not necessary to submit it to the Research Ethics Committee, following the guidelines of Resolution 510/2016 of the National Health Council. The analyses were conducted exclusively based on interactions with GAI models, whose responses are public and not linked to identifiable individuals.

7. Final Considerations

This study presented an approach to comparative usability evaluation, applying four different GAI models to the official websites of the COP30 and COP29 climate conferences. Based on the principles of ISO 9241-210, a standardized evaluation protocol was drawn up, using the same analysis command for all the tools. The responses were organized and classified according to the standard's criteria: Task Adequacy, Individualization Adequacy, User Expectations Conformity, and Controllability. The targeted categorical content analysis made it possible not only to systematize the qualitative data but also to compare the performance of the tools in terms of the sensitivity and depth of the evaluations generated.

Can AI Judge Usability? The results indicate that GAIs are effective in identifying structural usability problems, such as deep navigation layers, terminological inconsistency, lack of visual feedback, and the absence of control features like sticky headers or breadcrumb trails. These findings confirm the utility of GAIs in mapping general patterns and issues aligned with heuristic and ergonomic principles. However, critical limitations remain. The tools demonstrated a low capacity to assess technical accessibility criteria, such as color contrast ratios or screen reader compatibility, and produced inconsistent results in subjective dimensions, including user satisfaction and emotional impact. Moreover, disparities in output quality and interpretative depth across tools highlight the dependency of these systems on input quality and prompt design, reinforcing the importance of methodological rigor.

Beyond the analytical outcomes, the comparative application of multiple GAIs contributes to advancing methodological practices in HCI, particularly in large-scale and high-visibility scenarios like the COP websites. The triangulation of different GAI outputs allowed for the identification of convergent insights while also exposing divergence patterns that would likely be overlooked in a single-tool analysis. This

suggests that GAIs may serve as complementary instruments, especially for preliminary diagnostics and iterative design workflows, but not yet as standalone evaluators.

Finally, this study reaffirms that the integration of GAIs into usability assessment demands a critical perspective grounded in human-centered values. While automation can enhance efficiency and scalability, particularly in early stages of design validation, the interpretive, cultural, and experiential dimensions of usability still require human judgment. Thus, rather than replacing expert evaluation, GAI should be seen as part of a hybrid evaluative ecosystem that combines computational power with contextual sensibility.

7.1. Limitations and Threats to Validity

Despite its contribution, the study has limitations. Firstly, there was no external validation of the GAI responses by human experts or real users, which restricts the triangulation of results. Categorization, although systematic, was carried out manually, which can introduce interpretative biases. In addition, reliance on a single standardized prompt can influence the depth or focus of the responses generated. There is also a technical limitation related to the models' access to the websites, as observed with Gemini, which partially compromised the analysis. These limitations do not invalidate the findings but indicate that the results should be interpreted with caution and in their proper experimental context.

7.2. Future works

The next steps of this study aim to advance in two main directions. The first involves validating the results with human experts to assess the accuracy and completeness of the analyses produced by GAIs. Triangulating automated evaluations with assessments conducted by usability and accessibility professionals will help identify overlaps, gaps, and interpretive biases. The second direction focuses on expanding the evaluation criteria by incorporating complementary normative frameworks, such as ISO/IEC 25010, which outlines quality models for software products, and the Web Content Accessibility Guidelines (WCAG), which are important for assessing compliance with international digital inclusion standards.

Acknowledgements

We thank the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) for their support. This study used GAI tools in the empirical investigation stage. The responses analyzed were generated by the ChatGPT (OpenAI), DeepSeek, Gemini (Google), and Microsoft 365 Copilot platforms, which comprised the qualitative database for the usability analyses conducted. In addition, the tools DeepL Translate and Grammarly were used to support the translation and linguistic revision of the English version of this paper.

References

Alsadi, A. e Miller, D. (2023). Exploring the impact of artificial intelligence language model chatgpt on the user experience. *International Journal of Technology, Innovation and Management (IJTIM)*, 1:2023.

Araújo, M. F. B. d., Mota, M. P., e Seruffo, M. C. d. R. (2024). Combinando inteligência artificial generativa e inspeção humana: Uma análise da usabilidade do site da greenpeace brasil. In *Anais Estendidos do XXIII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC)*, pages 81–85. Sociedade Brasileira de Computação.

Bardin, L. (2011). Análise de conteúdo. ed. *Revista e Ampliada. São Paulo: Edições*, 70.

Bevan, N., Carter, J., e Harker, S. (2015). Iso 9241-11 revised: What have we learnt about usability since 1998? In *Human-Computer Interaction: Design and Evaluation: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part I 17*, pages 143–151. Springer.

Bleichner, A. e Hermansson, N. (2023). Investigating the usefulness of a generative ai when designing user interfaces. Master's thesis, Uppsala University.

Borges, J. M. e Araújo, R. D. (2024). Experiences and challenges of a redesign process with the support of an ai assistant on an educational platform. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*, IHC '24, New York, NY, USA. Association for Computing Machinery.

Brazilian Federal Government (2024). COP 30 in Brazil. `https://www.gov.br/planalto/pt-br/agenda-internacional/missoes-internacionais/cop28/cop-30-no-brasil`. Accessed: April 25, 2025.

Brooke, J. (1996). Sus: A 'quick and dirty' usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., e McClelland, I. L., editors, *Usability Evaluation in Industry*, pages 189–194. Taylor & Francis, London.

Castells, M. (2008). The new public sphere: Global civil society, communication networks, and global governance. In *The aNNalS of the american academy of Political and Social Science*, pages 78–93. Publishing Press.

Cyr, D., Head, M., e Ivanov, A. (2006). Design aesthetics leading to m-loyalty in mobile commerce. *Information & management*, 43(8):950–963.

Fischer, M. e Lanquillon, C. (2024). Evaluation of generative ai-assisted software design and engineering: A user-centered approach. In Degen, H. e Ntoa, S., editors, *Artificial Intelligence in HCI*, pages 31–47, Cham. Springer Nature Switzerland.

Hassenzahl, M. e Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, 25(2):91–97.

International Organization for Standardization (2010). ISO 9241-210: Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. `https://www.iso.org/standard/52075.html`. Accessed: 2025-04-24.

Ivory, M. Y. e Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Comput. Surv.*, 33(4):470–516.

Kuric, E., Demcak, P., Krajcovic, M., e Lang, J. (2025). Systematic literature review of automation and artificial intelligence in usability issue detection.

Li, D. (2024). Exploration of user experience design optimization for the campus library information management system. *Journal of Education, Humanities and Social Sciences*, 37:6–15.

Liu, F. (2021). International usability testing: Why you need it. `https:www.nngroup.com/articles/why-international-usability-testing/`. Accessed: 2025-04-24.

Miraz, M. H., Ali, M., e Excell, P. S. (2017). Multilingual website usability analysis based on an international user survey. *CoRR*, abs/1708.05085.

Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J. e Mack, R. L., editors, *Usability Inspection Methods*, pages 25–62. John Wiley & Sons, New York.

Nielsen, J. (1999). *Designing web usability: The practice of simplicity*. New riders publishing.

Norman Donald, A. (2013). *The design of everyday things*. MIT Press.

Oliveira, L. F. P. d. e Ferreira, S. B. L. (2022). Usabilidade e acessibilidade: Um estudo de caso com a plataforma instagram. `https://bsi.uniriotec.br/wp-content/uploads/sites/31/2022/04/202202LucasOliveira.pdf`. Trabalho de Conclusão de Curso, Universidade Federal do Estado do Rio de Janeiro (UNIRIO).

Pereira, R., Darin, T., e Silveira, M. S. (2024). Grandihc-br: Grand research challenges in human-computer interaction in brazil for 2025-2035. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*, pages 1–24.

Serra, L., Carvalho, L., Ferreira, L., Vaz, J., e Freire, A. (2015). Accessibility evaluation of e-government mobile applications in brazil. *Procedia Computer Science*, 67:348–357.

Shneiderman, B. e Plaisant, C. (2010). *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.

Vatrapu, R. e Pérez-Quiñones, M. A. (2004). Culture and international usability testing: The effects of culture in structured interviews. *CoRR*, cs.HC/0405045.