Journal on Interactive Systems, 2026, 17:1
ISSN: 2763-7719 • doi: 10.5753/jis.2026.7129

Research Paper

Human Perceptions Meet AI Analysis: A Usability Evaluation of Climate Conference Websites

Danilo T. Lima [Federal University of Pará | *danilo.teixeira@itec.ufpa.br*]
Alana M. Medeiros [Federal University of Pará | *alanamirandaufpa@gmail.com*]
Rita de Cássia R. Paulino [Federal University of Santa Catarina | *rcpauli@gmail.com*]
Marcos César R. Seruffo [Federal University of Pará | *seruffo@ufpa.br*]

Institute of Technology, Electrical Engineering Graduate Program, Federal University of Pará, Av. Augusto Corrêa, 01, Guamá University Campus, Belém, PA, 66075-110, Brazil.

Abstract. *Introduction:* Digital platforms play a central role in mediating major international events, serving as gateways for information, mobilization, and public engagement. Evaluating usability in heterogeneous, cross-cultural contexts such as the Conference of the Parties (COP) poses significant challenges. The emergence of Generative Artificial Intelligence (GAI) offers new opportunities for automated usability assessment, but its alignment with real user experience remains underexplored. *Objective:* This study critically compares GAI-generated usability evaluations against human perceptions of climate conference websites. GAIs assessed both COP29 and COP30 platforms, while human evaluations focused exclusively on COP30 during the event in Belém, Brazil. The objective is to examine convergences and divergences between automated and human-centered assessments. *Methods:* Four GAIs (ChatGPT, Gemini, DeepSeek, and Copilot) performed heuristic evaluations based on ISO 9241-210 criteria. GAI outputs were analyzed using Directed Categorical Content Analysis. These findings were complemented by an exploratory evaluation involving ten human participants ($n = 10$) with diverse profiles, whose perceptions were collected via an online questionnaire during COP30. Final analysis involved triangulation of GAI findings with human feedback. *Results:* Results revealed a notable divergence in overall assessment. Humans reported high satisfaction and usability for critical tasks, contrasting with GAIs, which penalized COP30 for structural deficiencies, particularly low Controllability. However, convergence occurred in Individualization Adequacy, the lowest-scoring criterion for both GAIs and humans, confirming systemic failures in adaptability and clarity. *Conclusion:* GAIs and users converge on structural issues but diverge on subjective and contextual evaluations. This highlights the value of hybrid models: GAIs provide systematic heuristic data, while human feedback grounds practical and sociocultural interpretations. Future work should include longitudinal evaluations, automated accessibility checkers, and hybrid workflows for robust assessment frameworks.

Keywords: Generative AI, Usability Evaluation, Human-Computer Interaction, Climate Conference Websites

Edited by: Maria Lúcia Bento Villela | **Received:** 07 December 2025 • **Accepted:** 17 August 2026 • **Published:** 31 August 2026

1 Introduction

The digital experience plays a decisive role in mediating between global audiences and major international events, functioning both as a gateway for information and as an environment for political mobilization and engagement. Digital platforms associated with the Conferences of the Parties (COP) on climate change, especially their official websites, have become strategic spaces for broadening the reach of agendas, facilitating understanding of the topics under debate, and supporting the participation of the diverse user profiles involved in the event.

The usability of these environments influences the capacity for navigation, interpretation of information, and public engagement, particularly when considering the sociocultural and technological heterogeneity that characterizes the global audience of these conferences [Castells, 2008; Oliveira and Ferreira, 2022].

In this scenario, evaluating the usability of major international event websites is increasingly necessary. The case of COP30, held in Belém, the capital of the state of Pará, in the Brazilian Amazon, highlights this relevance. Besides representing a political milestone for Brazil by placing the Amazon region back at the center of climate negotiations, the event also triggered expectations of high traffic, broad circulation of content, and a strong presence of diverse audiences [Brazilian Federal Government, 2024].

The inclusion of COP29, hosted in Azerbaijan, provides a geographical, structural, and informational contrast, allowing for a transversal investigation of usability across highly different cultural and political environments: one from a major South American nation focusing on biodiversity and traditional communities, and the other from a country in the Caucasus/Eurasian region, representing a distinct geopolitical and economic profile.

This pairing enables the study to observe how design choices, cultural responsiveness, and information architecture manifest across heterogeneous settings, offering insights that are not reducible to a single context. The visibility given to Amazonian communities, their practices, knowledge, and demands further amplified the role of digital platforms as mediators of discourses, experiences, and symbolic disputes.

Despite this growing importance, gaps are still evident in the literature concerning accessible, replicable, and automated methods that assist design and communication teams in carrying out agile and contextualized diagnoses of usability [Kuric *et al.*, 2025].

Starting from this gap, the first version of this study, "Can

AI Judge Usability? A Comparative Analysis of Generative Tools on Climate Conference Websites", published in the proceedings of the Brazilian Symposium on Human Factors in Computing Systems (IHC), presented a comparative analysis of the usability of the COP29 and COP30 websites based on the principles of the ISO 9241-210 standard, conducted by four Generative Artificial Intelligences (GAIs): ChatGPT, Gemini, DeepSeek, and Microsoft 365 Copilot, all activated by a single standardized prompt [Lima *et al*., 2025].

For the expanded version now presented, the investigation was broadened with the inclusion of a new dataset: the perceptions of different categories of COP30 participants during the days of the event, including researchers and students, journalists and communicators, civil society representatives, and members of governmental organizations. A total of ten participants $(n = 10)$ were included in this exploratory evaluation.

The incorporation of these accounts allowed for a comparison between the GAI evaluations of the COP30 website and real user experiences in the context of use, enriching the analysis and challenging the performance of the GAIs when confronted with situated, culturally marked interpretations influenced by the unique dynamics of a climate mega-event.

The decision to collect human evaluations exclusively from the COP30 website, rather than extending the participant study to COP29, was grounded in the study's objectives. Human data collection was conducted in situ during the COP30 event in Belém, Brazil, enabling the capture of situated, context-rich perceptions from users actively engaged with the platform for authentic, task-oriented purposes during the event itself.

This design aligns with established perspectives in usability research that emphasize the importance of ecological validity and real-world usage contexts in HCI evaluation [Tractinsky, 2018; Hornbæk and Hertzum, 2017], as opposed to remote evaluations that may abstract users from the conditions in which interaction actually occurs. The COP29 website was therefore not included in the human evaluation due to the absence of an equivalent field setting and the lack of opportunity for in-person data collection in that context.

Given this expanded scope, the guiding question for this version is now: "In what manner do GAIs and real users converge or diverge in the evaluation of the usability of the COP30 website? Additionally, how do GAI evaluations of COP29 and COP30 compare, and what do these findings indicate for Human-Computer Interaction practices?".

Based on this question, the objectives of the study were: (i) to present the usability evaluation of the COP29 and COP30 websites according to the principles of ISO 9241-210, using four GAIs; (ii) to collect, organize, and analyze the perceptions of distinct profiles of COP30 participants regarding the navigation experience; (iii) to compare the evaluations generated by the GAIs with the real user accounts, identifying approximations, tensions, gaps, and possible biases; and (iv) to discuss the implications of these results for HCI practices, especially concerning the use of GAIs in the evaluation of interfaces aimed at culturally diverse audiences.

By confronting automated evaluations with situated human perceptions, this work seeks to highlight both the potential and limitations of GAIs. The study's main scientific contribution is expected to be threefold. First, it aims to systematically identify where GAIs and human users converge, particularly on structural usability issues such as navigation depth, content fragmentation, and terminology inconsistencies, suggesting that GAIs may be effective at recognizing predictable, code-detectable design flaws; Second, it intends to reveal where they diverge, especially on subjective, contextual, and culturally marked dimensions such as satisfaction, emotional response, and the perceived relevance of design elements, indicating that GAIs may lack the situated grounding necessary for nuanced evaluation; Third, it seeks to expose dissonances among the GAIs themselves, such as potential technical failures and varying analytical biases.

This study is aligned with contemporary discussions on the impacts of artificial intelligence on HCI, particularly within the scope of Global Challenge GC6 of the GranDIHC-BR agenda, which addresses the ethical, paradigmatic, and diversity, equity, and inclusion dimensions in the relationship between AI and interaction [Duarte *et al*., 2024]. By confronting automated evaluations with situated human perceptions, this work emphasizes the need for contextual sensitivity and attention to the risks of reinforcing biases and exclusions in digital environments.

1.1 Ethical issues

The research was conducted following the ethical principles recommended by the Code of Conduct of the Brazilian Computer Society (SBC) and the general guidelines of COPE (Committee on Publication Ethics). This study is a direct extension and methodological application of the broader research project "Multimodal Evaluation of User Experience in News Websites and Web Systems: Integration of Methods and Data Visualization" which received a favorable ethical opinion from the Research Ethics Committee of the Federal University of Pará (CEP-UFPA) under CAAE number 65872122.0.0000.0018 and Parecer number 7.257.775, issued on November 29, 2024. The current investigation applies the previously approved methods for user experience evaluation to a new but thematically and methodologically aligned context (usability evaluation of climate conference websites), involving no additional ethical risks beyond those already assessed and approved.

During visits to the COP30 spaces, including the Green Zone and the Blue Zone, researchers approached potential participants only for a brief initial conversation, lasting approximately 2 to 3 minutes, in which they presented the study proposal, its objectives, and the form of participation. Only individuals who demonstrated spontaneous interest in collaborating received the link to the online questionnaire.

Data collection was carried out exclusively through a Google Forms questionnaire, where respondents had access to the Informed Consent Form (ICF). This document explained the purpose of the research, the absence of risks, the anonymous nature of the responses, the freedom to discontinue participation at any time, and the exclusively academic use of the data. No information that would allow for individual identification was recorded, and all responses were treated in an aggregated manner. Furthermore, methods and references were reviewed in an effort to reduce biases related to gender, class, race, ethnicity, and other social markers.

2 Theoretical Survey

This section brings together the main conceptual foundations that underpin this study. It addresses the importance of usability in global digital platforms, reviews traditional methods of usability evaluation, examines recent advances related to GAI, and identifies gaps that motivate the proposed comparative analysis.

2.1 Usability in Global Digital Platforms: Frameworks and Influences on International Engagement

In the context of major international events, such as the COPs of the United Nations Framework Convention on Climate Change (UNFCCC), digital platforms assume a central role in how global audiences access information, navigate complex agendas, and participate in political and environmental debates. These websites function beyond mere repositories of institutional content, serving as portals for registration, programs, and opportunities for engagement. Consequently, the usability of these interfaces directly influences the ability of diverse users, spread across distinct regions, languages, and technological conditions, to interact effectively with the climate negotiations [Nielsen, 1999].

The ISO 9241-210 standard [International Organization for Standardization, 2019] offers a framework for addressing usability in heterogeneous contexts, defining it as the extent to which specific users can achieve their objectives with effectiveness, efficiency, and satisfaction in a specified context of use [Bevan *et al*., 2015]. Its user-centered design principles, including iterative development cycles, feedback collection, and contextual analysis, are particularly relevant for platforms aimed at global audiences with variations in digital literacy, accessibility needs, and distinct cultural expectations [International Organization for Standardization, 2019]. Applied to the UNFCCC websites, these principles highlight the importance of clear information architecture, consistent navigation structures, and accessible paths for interacting with institutional and event content [Norman Donald, 2013].

The motivation for operationalizing usability through the three dimensions of Effectiveness, Efficiency, and Satisfaction is grounded in the task-oriented nature of the COP websites. These platforms are designed to support specific user goals: finding logistical information, accessing official documents, registering for events, and navigating complex schedules. For such platforms, the ISO 9241-210 framework offers a well-established, internationally recognized structure that directly addresses these task-based interactions [Bevan *et al*., 2015; International Organization for Standardization, 2019]. While dimensions such as emotional engagement and aesthetic experience are relevant [Hassenzahl and Tractinsky, 2006], the urgency and informational demands of a climate conference prioritize functional effectiveness and efficiency. The inclusion of Satisfaction, as defined by the ISO standard, captures the subjective dimension of the user experience without expanding the scope to broader constructs that would require different methodological approaches.

The choice of ISO 9241-210 as the primary analytical framework for this study warrants critical reflection. This standard was selected because it offers a well-established, internationally recognized structure for usability evaluation, encompassing the three core dimensions of Effectiveness, Efficiency, and Satisfaction [Bevan *et al*., 2015]. However, its applicability to heterogeneous cultural contexts, such as the COP30 website serving Amazonian communities, is not self-evident. As discussed in the GranDIHC-BR agenda (GC3 - Plurality and Decoloniality), usability heuristics originating from the Global North may impose design paradigms that do not align with the needs, practices, or epistemologies of communities in the Global South [de Oliveira *et al*., 2024; Lima and de Lima, 2025].

This concern is particularly relevant given that the COP30 website explicitly aims to engage Amazonian communities, traditional populations, and other historically marginalized groups. Rather than assuming that "better performance" according to Northern metrics is universally beneficial, researchers must consider whether such improvements could paradoxically create new dependencies or forms of exclusion for communities with limited infrastructure. Drawing on Amartya Sen's capability approach, development must be understood as an expansion of freedoms and choices for individuals and communities, not merely the optimization of technical metrics [Sen, 2000]. This critical perspective does not invalidate the use of ISO 9241-210, but rather situates it as one among multiple possible frameworks, whose applicability must be assessed in light of the specific cultural, infrastructural, and epistemological contexts in which it is deployed.

Beyond the normative scope, previous research demonstrates how usability directly influences participation on global platforms. Nielsen [1999] discusses that poor design, confusing menus, or inaccessible layouts can discourage engagement, especially among users from regions with limited infrastructure or fewer digital resources. Cultural preferences also shape the way international audiences interpret and navigate websites: for example, Cyr *et al*. [2006] shows that layout organization, visual hierarchy, use of colors, and location influence user satisfaction in multicultural contexts.

These cultural preferences operate at multiple levels. At a perceptual level, research has shown that users from different cultural backgrounds exhibit distinct patterns of visual attention, with some cultures favoring holistic, context-rich layouts while others prefer focused, linear structures [Nisbett, 2003]. At a cognitive level, cultural values such as individualism versus collectivism influence how users interpret navigation structures: users from individualistic cultures may prefer direct, task-oriented paths, whereas those from collectivist cultures may value exploratory browsing and contextual information [Marcus and Gould, 2000; Vatrapu and Pérez-Quiñones, 2004]. At a symbolic level, colors, icons, and metaphors carry culturally specific meanings that can either facilitate or hinder comprehension [Hofstede, 2001; Cyr *et al*., 2006]. These findings suggest that COP websites can benefit from culturally responsive design choices, multilingual support, and adaptive content presentation.

Additionally, studies on interface design for intercultural contexts emphasize the importance of visual clarity and flexible structures to accommodate different levels of digital literacy. Shneiderman and Plaisant [2010] argues that well-defined visual hierarchies and adaptable content formats help

reduce cognitive load, especially for users less familiar with technology or with limited connectivity. Complementarily, Hassenzahl and Tractinsky [2006] emphasizes the role of aesthetic quality and emotional engagement in the user experience, reinforcing the idea that platforms like the COPs, by balancing institutional seriousness with accessible design, can foster trust, understanding, and motivation for participation.

2.2 Challenges of Traditional Usability Methods in Global Contexts

Classic usability evaluation methods play a fundamental role in interface analysis, but their application in large-scale, multilingual, and highly complex systems reveals important limitations Campos *et al*. [2025]. Heuristic evaluation, as proposed by Nielsen [1994b], is based on generalist principles that, despite being effective in many contexts, may not fully capture the sociocultural nuances that shape the interaction of diverse users on international platforms. Research indicates that cultural and linguistic factors influence the perception of usability, suggesting that heuristics conceived in homogeneous contexts may overlook relevant problems in heterogeneous environments [Vatrapu and Pérez-Quiñones, 2004].

User testing, a central pillar of human-centered evaluation, becomes logistically burdensome in global scenarios. As highlighted by Liu [2021], recruiting participants in multiple countries, considering differences in connectivity, cultural expectations, and technological familiarity, demands considerable resources and can extend evaluation deadlines. For large-scale platforms with defined timelines, such as official COP event websites, these limitations sometimes make extensive participatory testing unfeasible.

The System Usability Scale (SUS) continues to be widely used due to its simplicity and reliability [Brooke, 1996]; however, evidence points to its limitations when applied to complex and multilingual platforms. For instance, Miraz *et al*. [2017] observes that SUS scores may fail to detect problems related to inconsistent translations, cultural inadequacies in content organization, or navigation disparities between versions in different languages de Carvalho *et al*. [2026]. More recent approaches, such as automated remote evaluations and interaction data mining, emerge as alternatives to overcome such limitations but lack validation in complex institutional contexts, such as the COP websites. Considered together, these challenges demonstrate the need for complementary evaluation strategies capable of dealing with cultural diversity, linguistic variability, and interface complexity, opening space for the emergent use of GAI.

2.3 Related Work: Potential and Limitations of GAI in Interface Evaluation

The automation of usability evaluations has been a recurring theme since the early 2000s, when pioneering research sought to systematize the process and reduce operational costs. The work by Ivory and Hearst [2001] organized existing methods into a taxonomy aimed at automation, highlighting its potential to streamline analyses. However, traditional approaches based on programmed heuristics face restrictions in highly complex scenarios, such as multilingual platforms, interfaces intended for global audiences, or systems with multiple user profiles. In recent years, the incorporation of GAIs has expanded these possibilities, offering more flexible and contextualized means to support usability evaluations.

Recent investigations explore how GAIs can assist in interface design and analysis. Fischer and Lanquillon [2024] examined the role of GAI in the software development cycle, identifying tasks where GAIs increase the efficiency and creative capacity of teams. Complementarily, Bleichner and Hermansson [2023] evaluated the use of stable diffusion models to produce interface prototypes, generating login screens from textual descriptions provided by users. The proposal was subjected to empirical testing, demonstrating the GAI's ability to transform textual commands into functional components.

The literature still lacks studies that systematically compare the performance of GAIs with consolidated usability evaluation methods. One of the few investigations of this type is presented by Araújo *et al*. [2024], who combined automated analysis with human inspection to evaluate the institutional website of Greenpeace Brazil[1], using ISO 9241-210 as a reference. The study demonstrates how hybrid analyses can increase the precision of the evaluation by integrating human and computational perspectives. Also relevant is the experience report by Borges and Araújo [2024], who analyze the use of tools like GPT-4 and Gemini 1.5 in the redesign of the educational platform Classroom eXperience, showing how these technologies offer useful textual diagnoses and improvement suggestions to the design process.

In addition to studies directly associated with the use of GAIs, other research contributes to understanding usability challenges in institutional contexts. Serra *et al*. [2015] analyzed mobile e-government applications in Brazil, identifying accessibility barriers based on formal guidelines. Although it does not involve GAIs, the investigation highlights the importance of systematic evaluations on governmental platforms, a concern that also guides the comparative analysis of the COP29 and COP30 websites conducted in this work.

In a more recent scenario, Alsadi and Miller [2023] investigated the impact of language models, especially ChatGPT, on the user experience. Their results indicate that the quality of the generated responses influences perceptions of trust, engagement, and relevance, reinforcing that GAIs rely heavily on context to produce useful analyses. In a complementary manner, Li [2024] discusses improvements in the user experience of a university library system, emphasizing that navigation and information organization processes must consider users' mental models and expectations, principles aligned with ISO 9241-210.

Despite these contributions, comparative studies in large-scale environments that consider cultural and linguistic differences among users are still scarce. In this sense, the present work expands the debate by applying multiple GAIs in the evaluation of the COP29 and COP30 websites, taking the ISO 9241-210 principles as a reference. The proposal allows observation of how different design decisions affect usability in a transversal and contextualized manner.

By engaging with limitations pointed out in previous studies, such as the rigidity of automated heuristics [Ivory and Hearst, 2001], the need for contextual sensitivity in AI-

[1] https://www.greenpeace.org/brasil/. Access on 26 August 2026.

supported evaluations [Alsadi and Miller, 2023], and accessibility obstacles in public platforms [Serra *et al.*, 2015], this investigation seeks to refine evaluation practices that are scalable, adaptable, and sensitive to user diversity. Thus, it contributes to the advancement of methodologies applicable to multilingual and heterogeneous environments [Li, 2024], reinforcing the importance of approaches that encompass methodological rigor and attention to cultural differences.

3 Research Methods

To develop this study, a multi-stage methodological procedure was adopted, structured to ensure the collection, organization, and systematic analysis of the data produced by the GAIs. The general flow of the methodological process is shown in Figure 1.

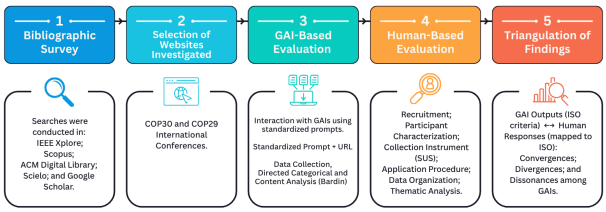

Figure 1. Methodological Path.

3.1 Bibliographic Survey

A structured bibliographic review was conducted to establish the broader theoretical context of this study and ensure the relevance and quality of the works supporting the analysis. The process included:

- (I) Delimitation of the investigation focus: The central question considered was: "In what manner do GAIs and real users converge or diverge in the evaluation of the usability of the COP30 website? Additionally, how do GAI evaluations of COP29 and COP30 compare, and what do these findings indicate for Human-Computer Interaction practices?".
- (II) Inclusion and exclusion criteria: Studies addressing GAI applications in usability, automatic evaluation methods, or topics related to Human-Computer Interaction were included. Incomplete texts or those not directly dealing with evaluative methods were disregarded.
- (III) Databases consulted: Searches were conducted in the IEEE Xplore (https://ieeexplore.ieee.org), Scopus (https://www.scopus.com), ACM Digital Library (https://dl.acm.org), Scielo (https://scielo.br), and Google Scholar (https://scholar.google.com/, using the descriptors 'Generative AI', 'Usability Evaluation', 'Automated Heuristic Evaluation', and 'Human-Computer Interaction'.
- (IV) Initial screening: Materials were selected by reading titles and abstracts, prioritizing relevance to the object of study.
- (V) Full reading and final selection: The selected texts were fully evaluated, observing methodological coherence with the research proposal.
- (VI) Qualitative synthesis: The results were organized to guide the methodological structure and subsequent discussions.

The complete technical report of the bibliographic survey, including search strings, screening results, and selected references, is publicly available in the Zenodo[2] repository.

3.2 Selection of Websites Investigated

For the analysis, the websites of the international conferences COP30[3] and COP29[4] were considered, both in their English versions. By standardizing the analysis on the English versions of sites, the study minimizes linguistic variables while maximizing the observation of design choices, cultural responsiveness, and information architecture across these heterogeneous settings. Figures 2 and 3 present the respective home pages, offering a visual overview of the analyzed interfaces.

Figure 2. Homepage of the COP30 Website.

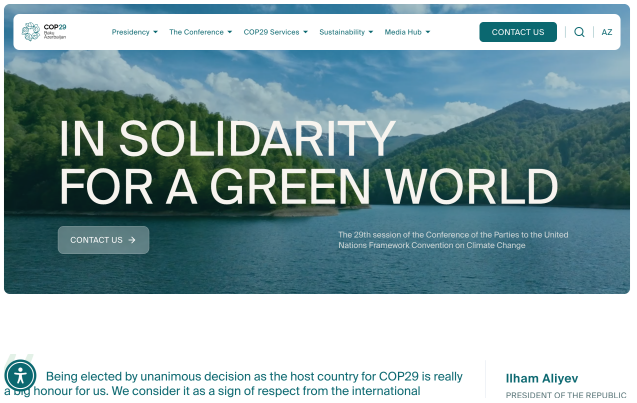

Figure 3. Homepage of the COP29 Website.

The choice was linked to the global relevance of these conferences, which represent the highest-level forum for international climate governance, but also to the opportunity to compare two platforms produced in distinct sociopolitical and geographical contexts.

The selection of the COP30 website is further justified by the historic significance of the event being held in Brazil, which prominently places the Amazon region at the center of

[2]https://zenodo.org/records/21138198
[3]https://cop30.br/en/
[4]https://cop29.az/en/

global climate negotiations. This context especially amplifies the website's interest for studies on digital accessibility, climate communication, and digital ecosystems aimed at the Amazon and its diverse, often marginalized, populations.

Conversely, the COP29 website, hosted in Azerbaijan, was included to provide an important point of geographical, structural, and informational contrast. This pairing allows for a transversal investigation of usability across highly different cultural and political environments: one from a major South American nation focusing on biodiversity and traditional communities, and the other from a country in the Caucasus/Eurasian region, representing a distinct geopolitical and economic profile.

The evaluation of the two websites by the GAIs was conducted on their desktop versions, between March 10 and 15, 2025, with a focus on navigation, access to official documents, and engagement features.

3.3 GAI-Based Evaluation

This subsection details the evaluation conducted with the four GAIs, corresponding to step 3 of the methodological path (Figure 1).

3.3.1 GAIs Used and Standard Prompt

Four GAIs were used to analyze the usability of the websites:

- (I) ChatGPT (OpenAI)[5] - version 4 turbo;
- (II) DeepSeek[6] - version 3;
- (III) Gemini[7] (Google) - version 2.0 flash;
- (IV) Microsoft 365 Copilot[8] - latest version (no specific version number).

All were used in the English version, with the think deeper option enabled; this decision was made to ensure comparability between the COP29 and COP30 platforms, as the Azerbaijani and Portuguese versions would introduce additional linguistic variables that could confound the usability evaluation.

To ensure comparability between analyses, the same standardized prompt was applied to all models, covering the three pillars of ISO 9241-210: effectiveness, efficiency, and satisfaction. 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;
> - 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."

A critical methodological clarification concerns how the GAIs processed the provided URLs. All four models were accessed through their standard web-based interfaces in text-only mode; none of them employed multimodal capabilities (such as screenshot capture, visual rendering, or pixel-based analysis) during the evaluation. When a URL is provided in the prompt, these models do not render the webpage visually, nor do they execute client-side scripts or capture screenshots. Instead, they access the publicly available HTML source code of the page, extracting its textual content, metadata, and structural elements (e.g., headings, paragraphs, lists, links, and alt attributes). The analysis is therefore based exclusively on the textual and structural representation of the website, rather than on a pixel-based or rendered visual inspection.

This distinction directly addresses the question of how the models evaluate visual aspects such as layout, contrast, and visual hierarchy. When the models report on visual hierarchy, contrast ratios, or button visibility, they are not perceiving these elements as a human user would. Instead, they infer them from the HTML structure and CSS declarations embedded in the source code. For example, a model may infer potential contrast issues by analyzing CSS color values and the structural relationship between interface elements. Likewise, observations regarding call-to-action buttons potentially blending into background images should be interpreted as code-based heuristic inferences rather than direct visual perception.

This mode of operation has both strengths and limitations; on one hand, it allows for systematic, scalable, and replicable analyses that are consistent across models and sessions. On the other hand, it means that the GAI evaluations are inherently theoretical and code-based, lacking the situated, perceptual, and contextual nuances of human visual inspection. Critically, the models cannot detect issues such as overlapping elements, rendering inconsistencies across browsers, or dynamic behaviors that depend on user interaction or screen resolution, as these require actual rendering and execution of client-side code, capabilities that were not available in the text-only mode used in this study. Therefore, the GAI-generated findings should be interpreted as heuristic diagnostics derived from the underlying code, rather than as direct visual assessments.

The instructions guided GAIs to produce more complete and uniform analyses, using the three central dimensions of ISO 9241-210 as a reference: Effectiveness (the website's ability to provide conference information and support audience engagement), Efficiency (the time and effort required to locate content and complete navigation tasks), and Satisfaction (users' subjective perception of ease of use, clarity of information, and overall experience) [Bevan *et al.*, 2015]. In addition, the GAIs were asked to make a direct compari-

[5] https://chatgpt.com/
[6] https://chat.deepseek.com/
[7] https://gemini.google.com/
[8] https://copilot.microsoft.com/

son between the COP30 and COP29 websites, identifying the strengths and weaknesses of each platform, as well as their relevant convergences and differences. Responses should also include specific recommendations for improvements to the COP30 website.

The standardized prompt was structured by the lead researcher, drawing upon the three central pillars of the ISO 9241-210 standard (effectiveness, Efficiency, and Satisfaction) as defined by [Bevan *et al*., 2015]. The prompt was designed to be comprehensive yet neutral, avoiding leading language that could bias the GAI responses. To ensure clarity, objectivity, and methodological soundness, the prompt underwent an internal review process involving two researchers with expertise in HCI and usability evaluation.

Their feedback led to refinements in the wording and structure, particularly regarding the specification of evaluation criteria and the request for concrete recommendations. While the prompt was not subjected to formal peer review prior to data collection, the iterative review process within the research team helped mitigate potential ambiguities and ensured alignment with the study's analytical objectives. This approach is consistent with practices in GAI-based research, which emphasize the importance of prompt engineering and validation through expert review [White *et al*., 2023; Meskó and Topol, 2023].

3.3.2 Data Collection

Interactions with each GAI were conducted in a controlled manner, always applying the same prompt and recording responses in full. This procedure allowed for replicability and control of variables that could interfere with the comparison between models.

3.3.3 Data Organization

The responses generated by the GAIs were organized based on a categorization method derived from the main criteria of ISO 9241-210. The criteria were: 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.

3.3.4 Directed Categorical Analysis

The content analysis followed the model proposed by Bardin [2011], using categories guided by previous theoretical references (in this case, the ISO standard itself). Selective coding was chosen to avoid typical repetitions in GAIs artificially inflating the frequency of certain elements. Only distinct and contextually relevant excerpts were considered; this procedure made it possible to examine qualitative differences in the way each GAI interpreted and evaluated the same interface elements.

Each usability finding generated from the GAI responses was assigned a unique identifier (ID) and systematically catalogued in a structured spreadsheet available in the Zenodo repository. This dataset, entitled "Categorization of AI-Generated Usability Findings by ISO 9241-210 Criteria", contains the structured mapping between the extracted usability findings and their corresponding, representing the final analytical layer derived from the GAI outputs.

3.4 Human-Based Evaluation

To complement the analyses carried out by the GAIs, an exploratory evaluation was conducted with 10 participants of different backgrounds present in the COP30 spaces, with the objective of identifying real user perceptions regarding the event website. This corresponds to step 4 of the methodological path (Figure 1). The procedure was structured as follows:

3.4.1 Recruitment

Recruitment took place in the public spaces of COP30 (Green Zone and Blue Zone) in Belém, Pará, Brazilian Amazon. Individuals were approached in person and invited to participate in the research. Only people who stated they had used the official COP30 website for informational or general navigation purposes were included.

The recruitment strategy aimed to maximize the sociocultural and technological diversity of participants, reflecting the heterogeneous profile of the conference audience. Therefore, no restrictive criteria were applied regarding age, nationality, or specific role/function at the time of data collection. Although the study sought to include participants from diverse backgrounds, the final sample should not be interpreted as fully representative of all groups attending COP30.

It is important to clarify why the study was not extended to online recruitment via social media platforms, despite the apparent feasibility of this approach. The decision to restrict data collection to in-person recruitment during the event was deliberate and grounded in methodological and epistemological considerations. First, the study's primary objective was to capture the situated user experience, perceptions shaped by the actual context of use during the event.

Participants recruited online, outside the physical and temporal boundaries of COP30, would likely report different impressions, influenced by their own devices, environments, and levels of engagement, potentially introducing confounding variables that could compromise the comparability with the GAI evaluations. Second, the research design prioritized ecological validity [Tractinsky, 2018].

By recruiting participants in situ, the study aimed to collect responses from individuals actively engaged with the platform for authentic, task-oriented purposes (e.g., checking schedules, finding logistics information, accessing documents), rather than from users who might access the site merely for the sake of the evaluation. This distinction is required, as usability perceptions are known to vary significantly between controlled testing environments and real-world usage scenarios [Hornbæk and Hertzum, 2017]. Third, the in-person approach enabled the research team to verify the essential inclusion criterion, that the participant had actually used the COP30 website, through direct conversation, reducing the risk of false claims that could compromise data quality. Online recruitment would have made such verification impractical.

Fourth, the conference environment itself was considered a key variable. The distractions, urgency, and emotional involvement characteristic of a major international event are not

easily replicable in an online setting. These contextual factors, while introducing limitations regarding the generalizability of the findings (see Section 6.1), were intentionally preserved as part of the phenomenon under investigation, aligning with the study's commitment to understanding usability as it is experienced rather than as it is reported in isolation. Finally, the exploratory nature of this investigation, with its emphasis on qualitative triangulation rather than statistical representativeness, did not require a large sample size. The priority was depth, contextual richness, and diversity of participant backgrounds.

Despite the high volume of individuals approached during the event days, several factors constrained the final sample size. First, although many attendees expressed initial interest in participating, a significant portion had not previously accessed the official COP30 website, a prerequisite for inclusion. Others, due to the fast-paced and demanding dynamics of the conference environment, completed the questionnaire in a cursory manner, providing responses that lacked analytical depth or were limited to the most superficial observations.

Additionally, some participants only answered the basic demographic and profile questions, leaving the core usability questions incomplete. These cases were systematically excluded from the final dataset to maintain analytical rigor. The participants who ultimately collaborated spanned a variety of profiles, including: Social or community movement representative; Governmental organization member; Journalist or communicator; Non-governmental organization (NGO) member; Researcher or student.

3.4.2 Participant Characterization

The ten participants who composed the final sample ($n = 10$) represented a range of demographic and professional backgrounds, including different genders, age groups, nationalities, and institutional affiliations. Although the sample included participants with varied profiles, it was not intended to be statistically representative of the broader COP30 audience.

Regarding gender identity, the sample consisted of 6 self-identified women and 4 self-identified men. In terms of race/color, based on self-declaration using the Brazilian Institute of Geography and Statistics (IBGE) classification, the participants included 4 White, 4 Brown (pardos), and 2 Black individuals. Concerning nationality, 7 participants were Brazilian, and 3 were foreign nationals (from different countries in Latin America and Europe), reinforcing the international character of the event. The age range varied from 24 to 51 years, with a mean age of approximately 35 years.

The data collection took place over eleven days during the COP30 event (November 10–21, 2025), with the research team actively recruiting participants on four of those days: two days in the Green Zone and two days in the Blue Zone. This distribution allowed the research team to capture perceptions from participants with different types of engagement with the conference.

The professional and institutional profiles of the participants included: 2 researchers or students (from public and private universities in Pará), 2 journalists or communicators (from independent media outlets), 2 social or community movement representatives (one from a women's movement and one from communities affected by dams), 2 governmental organization members (from the Brazilian federal government and the Pará state government), 1 civil society representative, and 1 NGO member (from an international organization). These profiles were included to increase the heterogeneity of the participant sample within the exploratory scope of the study. Given the exploratory nature of the study and the limited sample size, no comparative analyses between participant groups were intended or performed.

It is important to note that, despite the recruitment team's efforts, it was not possible to include participants from quilombola, riverside (ribeirinho), or indigenous communities during the data collection period. Consequently, the diversity achieved in the final sample reflects the participants who were available and willing to participate during the event, rather than the full sociocultural diversity represented at COP30. This limitation could reflect the logistical challenges and accessibility barriers faced by these populations in the context of a large-scale international event.

3.4.3 Collection Instrument

The survey was conducted using an online form, developed with Google Forms, containing both closed and open-ended questions organized into thematic blocks. The questionnaire was inspired by the SUS methodology [Brooke, 1996] and aligned with the three dimensions of ISO 9241-210: Effectiveness, Efficiency, and Satisfaction.

It is important to clarify that the instrument does not reproduce the original SUS in its entirety, nor does it employ its standardized 5-point Likert scale and 0–100 scoring formula. Instead, the SUS served as a conceptual and structural reference for organizing the usability dimensions investigated in this study. This approach is consistent with previous HCI research that has adapted the SUS framework to specific contexts without applying its full psychometric protocol, particularly in exploratory studies where the primary aim is qualitative triangulation rather than standardized benchmarking [Lewis, 2018; Bangor *et al.*, 2008].

The decision to adopt a simplified 3-point response scale (Disagree / Neutral / Agree), rather than the original 5-point Likert scale, was deliberate and grounded in practical and methodological considerations. First, the study was conducted in situ during the COP30 event, a setting characterized by urgency, environmental distractions, and high participant workload. A shorter, more intuitive scale reduces cognitive load and facilitates rapid completion, which was essential to maximize participation rates in this demanding context [Sauro and Lewis, 2011].

Second, the participant sample was diverse, including individuals with varying levels of digital literacy, educational background, and language proficiency (including non-native English speakers). Previous research suggests that simplified response formats can enhance comprehension and reduce measurement error in heterogeneous populations [Streiner *et al.*, 2016]. Third, the study's primary objective was qualitative triangulation, comparing human perceptions with GAI-generated evaluations, rather than producing a standardized SUS score or benchmarking against industry norms. The 3-point scale preserves the ordinal nature of the responses, enabling meaningful comparative analysis within the study's scope, while the loss of granularity is acceptable given the ex-

ploratory and triangulation-focused design [Carifio and Perla, 2007].

The complete instrument comprised 18 questions in total. However, for the purposes of the usability evaluation presented in this study, 10 core questions, those directly addressing the ISO 9241-210 criteria, were selected as the primary analytical focus. These 10 questions, presented in Table 2, correspond to the usability dimensions under investigation and were answered using a three-point scale (Disagree / Neutral / Agree), which was simplified to facilitate comprehension among the diverse audience expected at COP30, including non-native English speakers and individuals with varying levels of digital literacy.

The remaining eight questions served a complementary function; two questions characterized the participants' profiles (affiliation and purpose of access), three were open-ended questions capturing qualitative insights (what most helped, what most hindered, and suggestions for improvement), two addressed accessibility and inclusion (use of accessibility resources and perceived support) to characterize participants' perceptions regarding these aspects rather than to conduct a formal accessibility evaluation of the website, and one captured overall satisfaction on a 5-point Likert scale.

It is important to note that two response scales were intentionally employed for different analytical purposes. The ten core usability questions, inspired by the SUS framework, used a simplified three-point scale (Disagree / Neutral / Agree) to facilitate rapid completion and support comparisons with the GAI evaluations. In contrast, the overall satisfaction item employed a five-point Likert scale because its purpose was solely to capture participants' general satisfaction with the website and was not included in the triangulation analysis.

Before data collection, the questionnaire was reviewed by two researchers with expertise in HCI and usability evaluation to ensure clarity, relevance, and alignment with the study's objectives. Their feedback resulted in minor adjustments to the wording of certain items to enhance comprehensibility for the target audience. This review process, while not a formal peer review, helped ensure the instrument's face validity and appropriateness for the study's exploratory nature.

The complete questionnaire is available in the Zenodo repository in PDF format, providing a static record of the instrument as applied during the study.

3.4.4 Application Procedure

Following the agreement to participate, the link to the online questionnaire, hosted on Google Forms, was provided to the individual for completion on their own personal device. Participants were informed that they could complete the survey immediately or at a later time. Crucially, there was no researcher mediation or interference during the completion process, ensuring that the responses collected accurately reflected the participants' spontaneous and authentic perceptions of the website usability. The data collection was conducted between November 10 and November 21, 2025, within the Blue Zone and Green Zone spaces of COP30 in Belém, Pará.

While the active data collection was conducted via Google Forms, the questionnaire instrument made available in the Zenodo repository is provided in PDF format. This choice is intentional because the dynamic nature of online form platforms means that layouts, question order, or even the availability of the instrument could change over time. The PDF version serves as a timestamped, static snapshot of the instrument exactly as it was designed and presented to participants, ensuring methodological transparency and replicability. By depositing the PDF, other researchers can inspect, adapt, or replicate the questionnaire with full confidence in its fidelity to the version actually used in the study. This practice aligns with open science principles [Nosek *et al*., 2015].

3.4.5 Data Organization

The responses were grouped into three dimensions: Location of information, referring to the ease of searching for documents, schedules, and services; Structure and navigation, referring to the organization of sections and fluidity between pages; Comprehensibility, referring to textual clarity and use of technical terms. These dimensions were chosen to enable direct comparison with the ISO 9241-210 criteria adopted in the analysis of GAIs.

3.4.6 Analysis

It is important to clarify that the initial analytical strategy considered the use of directed content analysis, as employed in the evaluation of GAI outputs, following Bardin's framework. However, while content analysis is well-suited for categorizing data based on pre-existing theoretical frameworks, such as the ISO 9241-210 criteria, it proved less adequate for capturing the richness and diversity of the open-ended responses provided by the human participants.

The qualitative data from the participants were not limited to predefined categories; rather, they contained spontaneous narratives, contextual nuances, and emergent themes that were not fully anticipated by the ISO criteria alone. Thematic analysis was therefore adopted because it allows for a more inductive and flexible approach, enabling the identification of patterns that emerge directly from the data, while still accommodating deductive elements when relevant to the research questions [Braun and Clarke, 2006; Nowell *et al*., 2017].

A potential limitation of thematic analysis, particularly in the context of this study, is its reliance on the researcher's interpretative judgment, which may introduce subjectivity in the identification and labeling of themes. To mitigate this risk, the analysis was conducted iteratively, with regular discussions among the research team to review and refine the emerging themes, ensuring interpretive validity [Lincoln, 1980].

Additionally, the relatively small sample size limits the generalizability of the qualitative findings; however, the primary goal of this exploratory study was not statistical representativeness, but rather the triangulation of human perceptions with the systematically generated outputs of the GAIs. The findings from the thematic analysis were subsequently integrated into the discussion to directly contrast and triangulate the usability evaluations derived from the GAIs with the real-world experiences reported by users in the context of the event.

3.5 Triangulation of Findings

Following the independent analyses conducted by the GAIs (step 3) and the human participants (step 4), the final stage of the methodological procedure involved the systematic tri-

angulation of both datasets, corresponding to step 5 of the methodological path (Figure 1). This triangulation process was structured to address the study's core research question. The triangulation was conducted through three complementary steps:

i) The GAI outputs, which had been categorized according to the four ISO 9241-210 criteria (Task Adequacy, Individualization Adequacy, User Expectation Conformity, and Controllability), were organized into a structured framework. This framework served as the basis for comparison with the human evaluation data.

ii) The human participant responses, both quantitative (from the adapted SUS-inspired questions) and qualitative (from the open-ended questions), were mapped onto the same ISO criteria framework. The quantitative responses were normalized to a common scale (0–2), while the qualitative insights were analyzed to identify patterns and themes that could be compared with the GAI-generated findings. Although both datasets were normalized to facilitate visual comparison, they originate from fundamentally different evaluation processes. The participant data represent questionnaire responses, whereas the GAI data correspond to normalized frequencies of analytically distinct usability findings extracted from textual outputs. Therefore, the normalization supports comparative visualization of usability patterns rather than direct statistical equivalence between the two data sources.

iii) Convergences and divergences between the two datasets were systematically documented. Convergences were defined as instances where both GAIs and human participants identified similar usability issues or strengths. Divergences were defined as instances in which GAI evaluations and human perceptions differed significantly, particularly on subjective dimensions such as satisfaction, cultural relevance, and aesthetic judgment. Additionally, dissonances among the GAIs themselves were documented, highlighting inconsistencies in how different models evaluated the same interface elements.

4 Results

The results section presents the central findings derived from both the analyses carried out by the GAIs and the evaluations conducted by the human participants. The organization follows the criteria defined by ISO 9241-210, allowing the data to be examined across three complementary axes: (i) the websites' performance according to each usability criterion; (ii) a comparison between the COP30 and COP29 platforms; and (iii) a critical appraisal of the role of GAIs and human evaluators, highlighting convergences, divergences, and the limitations of each approach. The excerpts cited throughout the section include an identifier (ID) that makes it possible to trace their origin in the qualitative database. The complete spreadsheet containing all categories, codes, and analyzed records is made available on Zenodo repository.

4.1 Analysis by Criteria

Given the qualitative nature of the input, derived from extensive responses generated by four GAIs: ChatGPT, DeepSeek, Gemini, and Microsoft 365 Copilot, a Directed Categorical Content Analysis was implemented. The textual outputs were structured into three analytical blocks: 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 platforms. Each relevant text excerpt was systematically classified according to the ISO 9241-210 criteria and manually categorized, allowing for both quantitative (frequency of mentions) and qualitative (interpretation of patterns) analysis.

To mitigate the effects of redundancy and the overrepresentation of repeated concepts, a common characteristic of GAI output, a selective coding strategy based on the principles of targeted content analysis was adopted. Under this approach, only salient and contextually distinct passages were tallied for comparative frequency. Consequently, the frequencies reported here do not represent the raw totals of all phrases aligned with each ISO criterion, but rather reflect the number of unique and analytically relevant mentions per model, criterion, and website. This method ensures that the final data accurately reflect distinct analytical observations instead of counting paraphrased or semantically equivalent feedback from the same GAI.

These frequencies should not be interpreted as indicators of the severity or relative importance of usability issues. Instead, they provide an overview of how often distinct usability themes emerged in the GAI evaluations, serving as descriptive evidence to support the subsequent qualitative interpretation.

Figure 4 offers a comparative overview of the frequency of distinct usability findings reported by each GAI across the evaluated ISO 9241-210 criteria for the COP30 and COP29 websites. The radar chart visualization highlights both consistent patterns of assessment across the models and notable differences in how each GAI interpreted the websites' textual and structural representations according to the usability principles. The frequency values summarized in the figure were constructed from the synthesis of salient, non-redundant findings identified in the GAI responses, ensuring an analytical focus. These frequencies indicate the recurrence of analytically distinct usability themes rather than the relative severity or impact of the reported usability issues.

Figure 5 presents the same data in a grouped bar chart format, facilitating direct comparison of mention frequencies across GAIs for each usability criterion.

4.2 GAI-Based Comparison Between the COP30 and COP29 Websites

The GAI-based comparison between the COP30 and COP29 websites highlights both essential structural convergences and more prominent design divergences. Both platforms share a common foundational structure, providing important logistical information (such as visas, accommodation, and transport) in designated sections and relying on thematic navigation as a primary organizing principle. This structural similarity suggests adherence to established conventions for large-scale international event websites. As part of the evaluation process, it is worth noting that the response time of the GAI when processing the same prompt varied significantly: ChatGPT took 34 seconds, DeepSeek 36 seconds, Copilot 35 seconds, while Gemini required 4 minutes and 21 seconds. While not directly linked to the content of the analysis, this latency difference reflects system efficiency variations relevant to iterative usability evaluation workflows.

Following this structural overview, a qualitative analysis

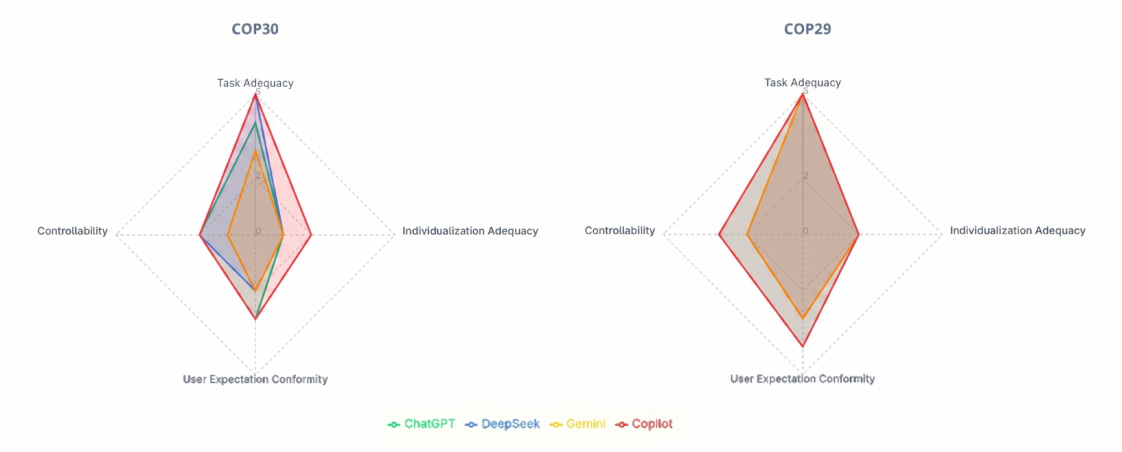

Figure 4. Graph of Frequency of Mentions by Usability Criterion.

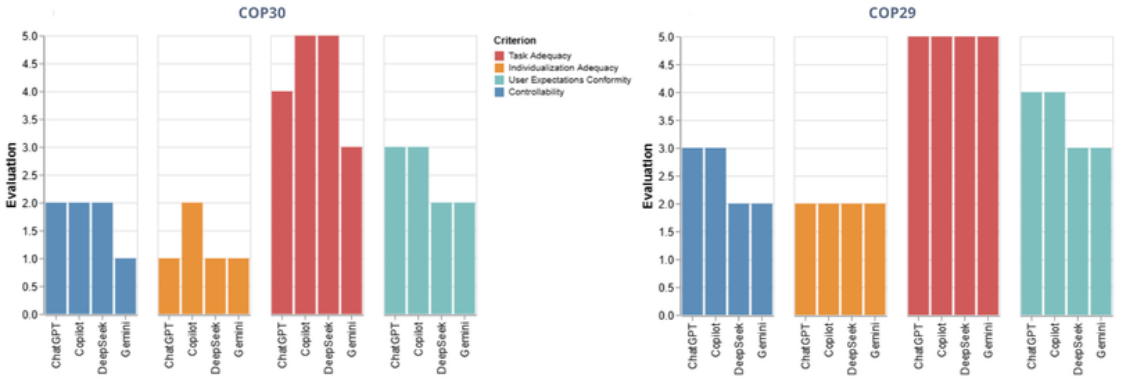

Figure 5. Frequency of Mentions by Usability Criterion and GAI Model.

was conducted using the GAIs' generated responses. Table 1 presents a summary of representative usability findings from these evaluations, contrasting key aspects across five usability dimensions: Navigation and structure; Visual design; Accessibility; Terminology and expectations; and User control. Each observation is substantiated by a traceable excerpt from the GAIs' outputs (indicated by their respective IDs). This comparative summary introduces the main distinctions identified and supports the detailed analysis that follows.

Differences in implementing core design principles led to distinct user experiences. For COP30, ChatGPT reported, *The site structure aligns well with the expectation of finding logistics information under a clear 'Logistics' section"* (ID: 9). At the same time, for COP29, it stated, *The homepage clearly directs users to key areas such as registration, schedule, and logistics."* (ID: 196). This contrast suggests that COP29 exhibited greater efficiency in its navigation architecture, utilizing shallower menus and more obvious main items. This approach, which aligns with consolidated HCI recommendations on navigation depth, was interpreted by the GAI as facilitating quicker access to critical information. Conversely, the GAIs suggested that COP30 placed a greater emphasis on visual identity elements, leading to an aesthetically richer experience.

However, this visual prioritization did not consistently translate into improvements in practical usability. Analysis suggests that this pursuit of visual differentiation may have compromised essential 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 choices compromised functional clarity.

Similarly, an exploratory divergence emerged concerning accessibility. ChatGPT identified the *The absence of an accessibility widget visible (e.g., text-to-speech, color adjustment tools)."* (ID: 12) for COP30, contrasting with the observation for COP29 that *Accessibility options such as screen reader optimization or high contrast modes are not clearly advertised"* (ID: 208). These observations should be interpreted as heuristic indicators derived from the GAI analyses rather than as validated accessibility assessments, suggesting potential differences in adherence to inclusive design principles that warrant dedicated accessibility evaluation.

4.3 Usability Evaluation with COP30 Participants

The usability evaluation of the COP30 website by human participants was conducted using a structured questionnaire, employing an adapted SUS methodology, to capture the real-world experience of use during the event days. The study included ten participants ($n = 10$) from different professional and institutional backgrounds (including researchers, journalists, and members of governmental and non-governmental organizations), ensuring a heterogeneous exploratory sample. Given the study design and sample size, the responses were analyzed collectively rather than by participant subgroup. The results, summarized in Figure 6, present the absolute fre-

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)

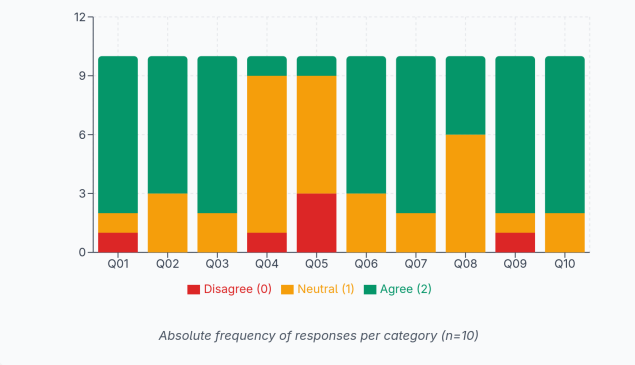

Figure 6. Response Distribution per Question.

quency of responses (Disagree, Neutral, Agree) for the ten questions (Table 2), allowing for an analysis of satisfaction levels and encountered difficulties.

The distributions presented in Figure 6 refer exclusively to the ten usability questions answered using the three-point response scale (Disagree / Neutral / Agree). Overall satisfaction was measured separately through a five-point Likert item, which was analyzed descriptively and was not included in the comparative triangulation with the GAI evaluations.

The participants who responded to the questionnaire appeared generally satisfied with the website's usability. In seven out of the ten questions, the "Agree (2)" category (represented by the green color) reached the maximum mark of 9 to 10 responses. This suggests that participants, broadly speaking, found the website easy to use, efficient, and intuitively navigable, which contrasts partially with some structural criticisms raised by the GAIs in the preceding analysis. The strengths perceived by human users appear to be linked to the effectiveness in locating essential information and general satisfaction with the interface, which is a positive indicator of usability. Although participants reported high overall satisfaction, the response distributions reveal greater disagreement regarding terminology consistency (Q04) and the availability of user support (Q05), indicating that positive overall evaluations coexisted with specific usability concerns.

This apparent coexistence of high overall satisfaction with localized usability concerns suggests that participants distinguished between their general experience with the website and specific interface characteristics. In the context of a live international event, successfully completing essential tasks may have contributed to positive overall evaluations despite the recognition of isolated usability issues.

However, these results also indicate notable points of friction, concentrated in questions Q02, Q04, and Q05, where disagreement and neutrality were more pronounced. Question Q08 also showed some neutrality, though to a lesser extent. Q05 (represented by the central bar with the largest portion in red and orange) is particularly relevant, showing the highest rate of negative responses (Disagree). If Q05 is interpreted as relating to design consistency, complexity, or clarity (aspects frequently criticized by the GAIs as "Visual Design" and "Terminology and Expectations"), the difficulty reported by human users corroborates the GAIs' concern that COP30's effort toward visual differentiation resulted in negative trade-offs in functionality. The consistent presence of Neutral and Disagree responses in these three questions indicates that, despite the generally positive responses, a portion of users encountered interaction barriers, suggesting that usability

Table 2. Usability Evaluation Questions

Item	Question
Q01	Were the information you were looking for easy to find?
Q02	Did the website help you complete your tasks in a simple way?
Q03	Do you consider the COP30 website to be organized in a logical and clear manner?
Q04	Does the website provide options that allow you to adapt the visual presentation?
Q05	Do you feel that the website considers different types of users?
Q06	Is the navigation similar to other websites you are familiar with?
Q07	Were the terms and section names understandable to you?
Q08	Did you find the website's visuals and text appropriate to the purpose of the event?
Q09	During navigation, did you feel in control of what you were doing?
Q10	When you made a mistake, was it easy to correct or recover?

deficiencies are situated and depend on the specific task or user expectation.

4.4 Triangulation of GAI and Human Evaluation Findings

The synthesis of the COP30 datasets, comprising the GAI evaluations and the human participant responses normalized to a common scale (0–2), is presented in Figure 7. This visualization facilitates the observation of convergences and divergences across the four ISO 9241-210 criteria: Task Adequacy, Individualization Adequacy, User Expectation Conformity, and Controllability. The quantitative evaluation profiles demonstrate a notable pattern of divergence in overall assessment. Accordingly, the radar chart should be interpreted as a visual comparison of usability evaluation patterns rather than as evidence of direct quantitative equivalence between human responses and GAI outputs. The quantitative evaluation profiles demonstrate a notable pattern of divergence in overall assessment.

The Human $(n = 10)$ profile (represented by the blue line) consistently occupies the largest area, indicating a generally higher perception of usability across all criteria when compared to the average assessment of the individual GAIs. This finding aligns with the majority of "Agree" responses from the human survey (Figure 6).

To map these dynamics qualitatively and understand the underlying reasons behind these alignments and discrepancies, we triangulated the GAI-generated heuristic findings with the situated accounts of human users. This relationship is synthesized in Figure 8, which categorizes the evaluation outputs into three analytical zones: Shared Usability Concerns (Convergences), Heuristic Rigor and Technical Vulnerabilities (GAI-Specific), and Situated Utility and Goal-Driven Leniency (Human-Specific).

As shown in the intersection of Figure 8, the most critical convergence point between GAI and human evaluations lies within Individualization Adequacy. This criterion registered the lowest composite score across both groups. Historically, GAIs have been recognized for their systematic ability to scan structural and code-based barriers. In our study, ChatGPT, DeepSeek, Gemini, and Copilot reached a high consensus when flagging the total absence of visual customization widgets (such as text resizing or high-contrast toggles), the lack of adaptive dark mode, and the inability to customize user-specific pathways.

This theoretical diagnosis of code-level deficiencies matches the empirical reality of users in the field. In the human questionnaire (Figure 6), Q04 ("Does the website provide options that allow you to adapt the visual presentation?") and Q05 ("Do you feel that the website considers different types of users?") gathered the highest concentrations of "Disagree" and "Neutral" responses. Specifically, 90% of participants answered Q04 with Neutral or Disagree, and Q05 registered the highest absolute disagreement rate in the study. This strong alignment validates the capability of GAIs to act as reliable, low-cost proxies for predicting inclusive design flaws and severe accessibility barriers before human testing takes place.

Additionally, both evaluations converged on structural navigation depth and information fragmentation. While GAIs repeatedly criticized "deep submenu structures" (e.g., DeepSeek, ID 23) and the lack of content prioritization on the homepage (e.g., Copilot, ID 136), journalists and governmental representatives in the field corroborated these issues. Human respondents explicitly reported friction when trying to "find some sections" or navigate a "barely visible menu", citing a lack of intuitive grouping for dynamic event schedules.

The most prominent divergences occur in the criteria of Controllability and User Expectation Conformity, revealing a fundamental paradigm shift between a code-based audit and a lived user experience. As visualized in the left zone of Figure 8, GAIs strictly penalized the COP30 website's Controllability due to the formal absence of elements such as "Back to Top" buttons, breadcrumb navigation trails, or sticky navigation headers on long content pages. For the automated models, these omissions represented a significant loss of user autonomy and navigation control. However, the human data directly contradicts this negative outlook. In Q09 ("During navigation, did you feel in control of what you were doing?"), 8 out of 10 human participants agreed that they felt fully in control. In practice, human users easily bypassed these missing design elements by scrolling manually, demonstrating that automated heuristic tools tend to over-penalize theoretical design flaws while underestimating human adaptability.

This divergence is further explained by the user's focus on utility over aesthetic consistency. While GAIs heavily critiqued visual distractions, low background contrast on call-to-action buttons, and terminology inconsistencies like the coexistence of "Programme" and "Agenda" (ChatGPT, ID 6), human users remained highly satisfied. This interpretation

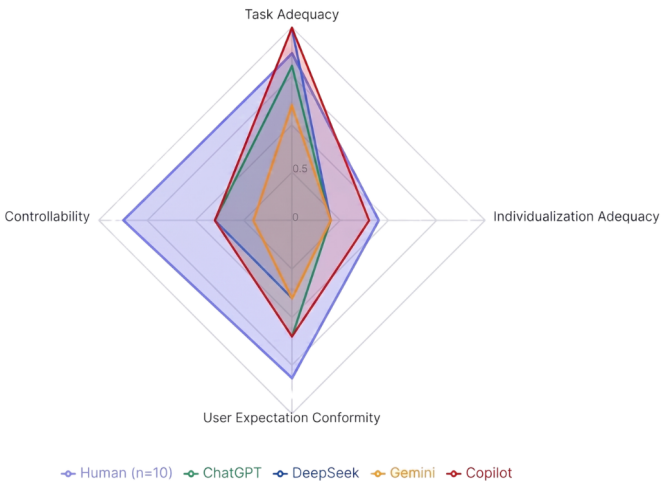

Figure 7. Multidimensional Evaluation Profile - Human vs. GAIs.

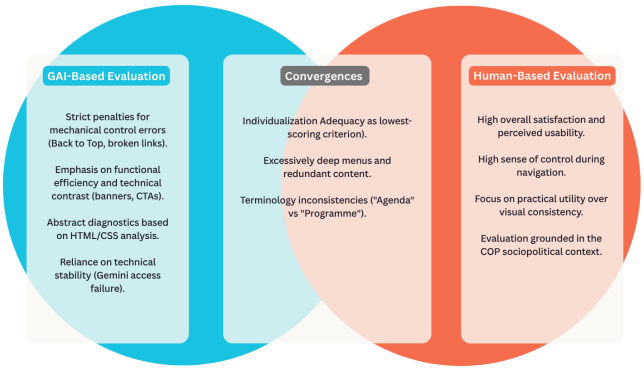

Figure 8. GAIs and Humans Agree on Structural Flaws, but Disagree on Satisfaction and Context.

is further supported by the separate overall satisfaction item, measured on a five-point Likert scale, in which participants predominantly assigned ratings of 4 or 5.

These findings suggest that overall satisfaction and specific usability judgments should not be interpreted as contradictory outcomes. Rather, they reflect different dimensions of user experience, in which participants acknowledged localized interface limitations while maintaining positive evaluations of the website's overall usefulness in supporting their goals during the event. Immersed in the demanding and high-stakes environment of a live international summit, participants prioritized the effectiveness of completing critical real-world tasks, such as finding logistics, schedules, and accreditation links, over formal interface consistency.

These findings highlight the boundaries and unique strengths of each evaluator type. On one hand, GAIs provide an exhaustive, replicable, and highly standardized checklist of predictable, code-detectable issues (like alt-text coverage, CSS contrast, or nested elements). However, they lack the emotional grounding, cultural context, and adaptive capability that humans possess. For example, the Gemini initially experienced an access failure before successfully completing the evaluation after the procedure was repeated. This episode illustrates the operational instability that may affect GAI-based evaluation workflows. On the other hand, non-expert human users provide the indispensable ecological validity required to understand how usability issues translate into actual experience.

These complementary profiles argue against relying on GAI-only evaluations for public platforms aimed at highly heterogeneous, global audiences. Instead, these findings suggest that a hybrid evaluation workflow could be explored in future research. In such a model, GAIs might perform the preliminary, low-cost structural sweep to flag potential accessibility and architectural inconsistencies, allowing human-centered testing to focus on validating these issues and exploring the situated, culturally marked dimensions of user engagement.

5 Discussions

The findings of this study reveal a complex, multi-layered landscape in which GAI analyses and human user experiences intersect, diverge, and occasionally challenge one another. By incorporating the evaluations of COP30 participants alongside the outputs of four GAIs, this investigation provides a more comprehensive understanding of how usability is perceived and judged when automated interpretations meet situated human experience. Overall, the results indicate that GAIs are competent in identifying structural issues and recurring design flaws, yet they are limited in capturing the contextual, emotional, and cultural nuances that strongly influence real user perceptions, particularly in global, politically charged environments such as the COP conferences.

A first major insight concerns the convergence between GAIs and human participants regarding fundamental structural usability problems. Both groups pointed to navigation depth, the fragmentation of content, and inconsistencies in terminology as recurrent weaknesses of the COP30 website. GAIs identified these patterns through analytical consistency across models, with ChatGPT, DeepSeek, and Copilot repeatedly highlighting "deep submenu structures", "multi-step access paths", and "terminology inconsistencies". Human users, while generally more positive, also expressed difficul-

ties in locating specific information (Q02, Q04, and Q05).

This overlap suggests that GAIs are highly effective at recognizing issues that interfere with task execution, namely, predictable errors related to information architecture and the predictability of interface elements. This confirms the framework of Bevan *et al*. [2015] and Norman Donald [2013]: usability principles remain valuable for identifying structural failures that affect task completion across all user groups. However, the convergence also reveals a limitation shared by both approaches, the difficulty in addressing the deeper cultural and contextual dimensions of usability, a concern raised by Nielsen [1994a] and Vatrapu and Pérez-Quiñones [2004].

However, the comparison also highlights critical areas of divergence, defining the boundaries of GAI evaluation. GAIs tended to heavily penalize elements based on aesthetic tensions, such as banner prominence, visual clutter, and insufficient contrast ratios, often classifying these as flaws. Conversely, human participants demonstrated a tendency to overlook these attributes in favor of the website's overall clarity and efficacy in completing their primary goals (utility). This contrast reflects a core GAI limitation: their assessments disproportionately rely on textual and structural surface-level patterns rather than lived interaction dynamics. The high score given by humans to Controllability directly contradicts the GAI's low rating, signaling that GAIs overestimate the impact of heuristic inconsistencies while underestimating the user's focus on goal prioritization.

The dissonance among the GAIs themselves, as summarized in Table 3, further illustrates this limitation. For the same visual hierarchy elements, ChatGPT noted "inconsistent font sizing" while DeepSeek highlighted that "homepage banners may distract from calls to action" and Gemini observed that the site "relies on large banners, pushing content below fold." Regarding navigation structure, ChatGPT considered the menus "intuitive," whereas DeepSeek criticized "submenus [that] require multiple clicks" and Gemini noted "vague labeling." These inconsistencies reveal that GAI evaluations are not only divergent from human perceptions but also internally inconsistent across models, each GAI interprets the same HTML/CSS code through its own analytical lens, producing different diagnoses for the same interface elements. This finding extends the work of Alsadi and Miller [2023], who demonstrated that generative models rely heavily on context to produce useful analyses, by showing that even when the context (the same prompt and the same URL) is held constant, different models produce divergent evaluations. This suggests that the variability is not merely a matter of contextual dependency but also of model architecture, training data, and analytical heuristics embedded in each GAI.

A second important tension emerges regarding contextual and cultural sensitivity. Participants representing various sectors (e.g., NGO, government) expressed concerns about the visibility and clarity of information regarding event logistics and politically relevant themes. GAIs, operating independently of the COP30's sociopolitical context, were limited in interpreting these expectations, focusing instead on generic interface principles. This finding reinforces the argument that GAIs lack the contextual grounding necessary for nuanced evaluation in politically charged settings, a limitation that directly responds to the concerns raised by Alsadi and Miller [2023] about the dependency of GAIs on context for producing useful analyses. Furthermore, the study confirmed a technical vulnerability: Gemini's failure to access the COP30 website demonstrates that automated evaluations are strongly dependent on external system stability. Unlike human users, who adapt and attempt alternative routes when facing errors, GAIs lack the autonomy to compensate for such technical barriers, raising serious concerns about the reliability of GAI-only evaluations in large-scale remote audits.

These concerns point toward the potential value of a hybrid evaluation model that could delineate which tasks might be effectively delegated to GAIs and which would continue to require human mediation. Based on the convergences and divergences identified in our findings, we propose the following division of responsibilities: Tasks suitable for automation include initial heuristic screeningk, such as identifying structural issues like navigation depth, content fragmentation, and terminology inconsistencies, which are predictable and code-detectable; systematic comparison across multiple interfaces, as demonstrated in the COP29 vs. COP30 analysis; generation of structured reports aligned with established standards; and flagging of potential accessibility violations based on HTML/CSS analysis, such as missing alt texts or insufficient contrast ratios.

Conversely, tasks that require human mediation include the interpretation of situated, context-dependent usability issues. GAI lacks the embodied experience and cultural grounding to assess how design choices affect real users in specific contexts, as evidenced by the divergence between GAI and human evaluations in dimensions such as satisfaction and cultural relevance. Human evaluators are also needed to validate and prioritize GAI-identified issues, distinguishing between heuristic violations that are genuinely problematic and those that are merely theoretical inconsistencies.

Furthermore, engaging with marginalized communities, such as Amazonian populations, traditional groups, and other historically marginalized peoples, requires participatory methods that center these communities as active agents rather than passive subjects of automated diagnostics. Finally, ethical deliberation and value-sensitive design cannot be reduced to algorithmic optimization, as decisions about what constitutes "better usability" in heterogeneous contexts involve value judgments that must consider local perspectives and capabilities [Sen, 2000].

In practical terms, one possible instantiation of this hybrid workflow could proceed iteratively. GAIs could conduct an initial automated sweep, producing a structured report with heuristic findings and flagged issues. Human evaluators would then review this report, validating or discarding findings based on contextual knowledge and direct engagement with users. For issues identified by GAIs but not supported by humans, the team conducts further investigation through targeted user testing or participatory workshops. Conversely, for issues flagged by humans but missed by GAIs, the team refines the prompts or develops new analytical criteria to improve the GAI's performance in future iterations.Such a process, however, would require empirical validation in future studies before it could be recommended with confidence.

Returning to the theoretical pillars that structured this

Table 3. 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)

investigation, some contributions to the existing literature emerge. Regarding usability in global digital platforms, the study provides empirical evidence that extends the work of Cyr *et al*. [2006], Nisbett [2003], Marcus and Gould [2000], and Hofstede [2001]. The divergence between GAI and humans on subjective dimensions confirms that cultural preferences, whether visual, cognitive, or symbolic, are not captured by universal heuristics. The convergence on Individualization Adequacy, however, reveals that even when GAIs and humans agree, the agreement is on a deficiency, pointing to a shared blind spot in both approaches regarding cultural adaptation, a concern raised by Shneiderman and Plaisant [2010] and Hassenzahl and Tractinsky [2006].

Concerning the challenges of traditional usability methods, the study responds to Liu [2021] observation about the logistical burdens of international recruitment by demonstrating the value of in-situ data collection during major events. The adaptation of the SUS instrument to a 3-point scale responds to Miraz *et al*. [2017] critique that standardized instruments may fail in multilingual contexts, offering a pragmatic compromise for exploratory studies. However, the adaptation also introduces tensions: consistent with Nielsen [1994a] heuristic evaluation framework, the simplified instrument captured broad tendencies but lacked granularity to detect subtle cultural mismatches, confirming Miraz *et al*. [2017] concerns while demonstrating that methodological flexibility can enhance participation from diverse audiences.

Regarding the potential and limitations of GAIs, the study extends Ivory and Hearst [2001] taxonomy by demonstrating that automation remains valuable for structural issues but faces restrictions in complex, multicultural scenarios. The divergence between GAI and human evaluations substantiates Alsadi and Miller [2023] argument that GAIs lack contextual grounding. The dissonance among the GAIs themselves, as illustrated in Table 3, adds a further layer: even when the same prompt and URL are provided, different models produce divergent diagnoses for the same interface elements. This suggests that the variability is not merely a matter of contextual dependency but also of model architecture, training data, and embedded analytical heuristics. The hybrid workflow we propose extends Araújo *et al*. [2024] demonstration of hybrid analysis by specifying a clear division of responsibilities, offering a practical framework for integrating computational consistency with human diversity and sensitivity.

Beyond these specific contributions, the study situates the COP websites as a lens for understanding a broader challenge: the communicative potential between automated evaluation systems and non-technical human evaluators. The participants (journalists, civil society representatives, community members, and researchers) were not usability experts, yet their perceptions were essential for assessing whether the platform served its intended purpose. This raises questions that extend beyond the COP context: How can automated tools be designed to complement, rather than replace, the situated knowledge of non-expert users? How can evaluation frameworks bridge the gap between technical optimization and meaningful user experience? The COP websites, in this context, serve as a case study for the increasing reliance on AI systems in public-facing digital platforms, where the stakes involve not only user satisfaction but also political engagement, cultural representation, and social inclusion.

The GranDIHC-BR agenda provides a framework for understanding the broader implications of this study through two interconnected Grand Challenges. GC3 - Plurality and Decoloniality emphasizes that HCI research must critically examine whether its methods serve or marginalize communities in the Global South [de Oliveira *et al*., 2024; Lima and de Lima, 2025]. The findings substantiate this concern: the divergence between GAIs and human participants regarding satisfaction and cultural relevance suggests that automated evaluations risk imposing Northern usability standards that may not align with the needs of diverse audiences. This is par-

ticularly significant given that the COP30 website explicitly aims to engage Amazonian communities, traditional populations, and other historically marginalized groups. Rather than assuming that "better performance" according to Northern metrics is universally beneficial, our results suggest that such improvements must be assessed in light of the specific cultural, infrastructural, and epistemological contexts in which they are deployed, a perspective aligned with Sen [2000] Amartya Sen's capability approach.

Complementing this decolonial perspective, GC6 – Implications of AI in HCI addresses the ethical, paradigmatic, and diversity, equity, and inclusion dimensions in the relationship between AI and interaction [Duarte *et al*., 2024]. Our findings contribute to this agenda in three ways. The dissonance among GAIs, each producing different diagnoses for the same interface elements (Table 3), raises ethical concerns about the reliability and transparency of automated evaluations. If different GAIs provide conflicting usability assessments, designers face uncertainty about which diagnosis to trust, with direct implications for equity and inclusion: automated evaluations may systematically favor certain usability paradigms over others, potentially marginalizing users whose preferences are not captured by the models' training data.

Furthermore, the divergence between GAI and human evaluations on subjective dimensions suggests that current models are not yet equipped to address the diversity of user perspectives that GC6 calls for, reinforcing the need to examine AI systems critically, not simply as technical tools but as sociotechnical systems embedded in power relations, cultural assumptions, and normative standards. The hybrid workflow we propose offers a practical response to these challenges: by reserving human mediation for tasks requiring contextual judgment, cultural sensitivity, and ethical deliberation, the model ensures that automated tools complement, rather than replace, the situated understanding of human evaluators.

The findings also invite a reflection on how cultural preferences shape the interpretation of layouts in multicultural contexts. As noted in the theoretical framework, users from different cultural backgrounds exhibit distinct patterns of visual attention: some cultures favor holistic, context-rich layouts, while others prefer focused, linear structures [Nisbett, 2003]. This distinction is particularly relevant for the COP30 website, which serves a global audience with diverse cognitive and perceptual expectations.

The divergence between GAIs and human participants, where GAIs penalized aesthetic elements that humans overlooked, suggests that automated evaluations, lacking cultural grounding, cannot account for these variations. Instead, they apply universal heuristics that may misinterpret culturally specific design choices as flaws. For example, a layout that appears "cluttered" to a user from a culture favoring minimalism may be perceived as "rich and informative" to a user from a culture that values contextual abundance. This has direct implications for the COP30 website: design decisions that prioritize visual identity and Amazonian symbolism may be misjudged by automated tools, while being positively received by the very communities the event aims to engage.

Likewise, the study's findings suggest that cultural preferences are not static but are mediated by the urgency and informational demands of the context. Participants in this study, immersed in the fast-paced environment of a climate conference, prioritized task completion and utility over aesthetic consistency. This contrasts with findings from studies conducted in less time-sensitive settings, where visual appeal and emotional engagement play a more prominent role [Hassenzahl and Tractinsky, 2006].

This suggests that cultural preferences are not merely about nationality or region, but also about the situational context in which users interact with a platform. For the COP30 website, this implies that usability evaluations, whether automated or human, must consider not only the cultural backgrounds of users but also the specific tasks and pressures they face during the event. Future research should explore how cultural preferences shift across different phases of an event (e.g., pre-event, during, and post-event) and how design can adapt to these varying expectations, rather than assuming a one-size-fits-all approach.

Taken together, these findings respond to the central question of the study: GAIs and real users converge on structural usability issues but diverge substantially in subjective and contextual evaluations. This divergence underscores the necessity of hybrid evaluation models in HCI. The GAI output provides efficient, systematic data for heuristic alignment, while human feedback provides the indispensable grounding for interpreting the practical and socio-cultural implications of usability challenges. Ultimately, this study suggests that GAIs and human participants provide complementary forms of evidence, supporting analytical strategies that integrate computational consistency with human diversity, contextual sensitivity, and situated perception in usability evaluation.

6 Final Considerations

This study combined GAI-based evaluations of both COP30 and COP29 websites with human participant assessments focused exclusively on the COP30 platform, collected during the event in Belém. By triangulating insights from four GAIs (ChatGPT, DeepSeek, Gemini, and Copilot) with real user perceptions from diverse participant profiles, the research offers a comprehensive view of how automated models and human evaluators interpret usability criteria grounded in ISO 9241-210.

The findings indicate that, although the evaluated GAIs were capable of producing structured, consistent, and prompt-responsive analyses, they did not fully approximate the contextual richness, cultural awareness, and situational reasoning demonstrated by users interacting with the COP30 platform during the event. This combined approach contributes to the methodological discussion by suggesting that, in contexts such as the one investigated, GAIs are better suited to complement rather than replace human-centered usability evaluations.

From a theoretical perspective, the results support the value of integrating automated tools into early-stage usability diagnostics while maintaining human oversight in interpretive, emotional, and context-dependent evaluations. The GAIs showed strong convergence when identifying structural issues such as navigation depth, content fragmentation, and inconsistent visual hierarchy. Human participants, however, emphasized experiential dimensions such as clarity of information, task completion, and perceived relevance, factors

shaped by their roles at the conference. This divergence confirms that effective usability evaluation, especially in highly heterogeneous contexts, requires hybrid models that leverage the scale of automated inspection without losing the situated knowledge produced by real users.

The comparison between COP30 and COP29, conducted exclusively through the GAI analysis, further illustrates the limitations of automation. GAIs consistently rated COP29 higher in structural efficiency while penalizing COP30 for visual choices that compromised functional clarity. This finding underscores the trade-offs between aesthetic identity and practical usability, a tension that automated tools are not equipped to interpret without human contextual grounding.

The hybrid workflow tentatively outlined in this study suggests a possible path forward, but should be viewed as a hypothesis rather than a validated recommendation. In such a model, GAIs could perform an initial structural sweep across platforms, flagging code-detectable issues, while human evaluators would validate, contextualize, and prioritize these findings. This proposal is offered as a conceptual stance that acknowledges the value of automation while insisting on the irreplaceable role of human judgment.

The study engages with the GranDIHC-BR agenda through two interconnected Grand Challenges; The first, GC3 - Plurality and Decoloniality, is addressed through the critical examination of Northern usability standards and their applicability to Amazonian communities. The divergence between GAIs and human participants regarding satisfaction and cultural relevance suggests that automated evaluations risk imposing universal heuristics that may not align with the needs of diverse audiences. The second, GC6 - Implications of AI in HCI, is engaged through the ethical and reliability concerns raised by dissonances among GAIs.

The inconsistencies in how different GAIs evaluated the websites through their textual and structural representations highlight the need for transparency and critical assessment of automated tools before they are adopted in high-stakes, public-facing platforms. By reserving human mediation for tasks requiring contextual judgment, cultural sensitivity, and ethical deliberation, the hybrid model responds to GC6's call for examining AI not only as a technical tool but also as a sociotechnical system embedded in power relations, cultural assumptions, and normative standards.

6.1 Limitations and Threats to Validity

Despite offering novel insights, the study presents limitations that should be considered when interpreting the results. First, the GAI evaluations relied entirely on external website access, making them vulnerable to technical failures. The Gemini model's inability to access the COP30 website illustrates a threat to internal validity: GAI outputs may be incomplete or misleading when access is unstable or blocked, producing false negatives or truncated analyses.

Second, human participant data were collected in situ during the COP30 event, a setting characterized by urgency, environmental distractions, and emotional involvement. These contextual pressures may have influenced participants' perceptions, potentially leading to more lenient evaluations due to the immediacy of their information needs. Rather than representing a threat to ecological validity, these contextual conditions reflect the authentic environment in which the website was intended to be used. However, they may limit the external validity and generalizability of the findings, as participants' perceptions were shaped by the specific demands and circumstances of a live international event. Additionally, the study employed a small sample size ($n = 10$) for the human evaluation. While this limits statistical generalizability, it is consistent with exploratory and qualitative research traditions. Nielsen [1994a] demonstrated that usability testing with as few as five participants can reveal approximately 85% of usability problems in an interface, with diminishing returns beyond this number. Faulkner [2003] further showed that small samples can be effective for qualitative insights, particularly when the objective is to identify the nature, rather than the prevalence, of usability issues. In this exploratory context, the sample size is justified by the study's analytical aim: to establish a dialogue between human perceptions and the systematically generated outputs of the GAIs, rather than to produce statistically generalizable data. The ten participants provided sufficient diversity across different backgrounds (researchers, journalists, governmental and non-governmental representatives) to enable meaningful triangulation with the GAIs' structured findings.

It is also important to clarify the exploratory nature of the human evaluation data and the limits of their interpretability. Although the results from the adapted SUS, inspired questionnaire are presented using quantitative descriptors, such as frequency of responses, normalization, and comparison with the average scores of the GAIs, these data do not possess inferential statistical validity. The small sample size ($n = 10$) and the adaptation of the SUS instrument to a simplified 3-point scale preclude the calculation of standardized SUS scores (0–100) and any form of statistical generalization to a broader population. Instead, the quantitative treatment of the human responses serves a purely descriptive and illustrative purpose: to identify emerging patterns, highlight convergent and divergent trends between human and GAI evaluations, and support the qualitative triangulation that constitutes the core of the study's analytical contribution. As such, the numerical summaries should be interpreted as qualitative indicators of tendency, not as statistically significant measurements. This distinction is critical to avoid overinterpreting the findings and to correctly situate the study within its exploratory, hypothesis-generating scope [Braun and Clarke, 2006; Sandelowski, 2000].

Third, accessibility was treated as a secondary analytical dimension that emerged from the GAI evaluations rather than as a predefined objective of the study. Consequently, the study did not include accessibility as a primary evaluation criterion. While the GAIs flagged some accessibility-related issues (e.g., contrast ratios, missing alt texts, and the absence of accessibility widgets), these analyses were derived from HTML/CSS inference rather than from actual assistive technology testing or user-centered accessibility evaluation. As noted in the findings, the GAIs identified the absence of accessibility features as a shared limitation of the COP30 website, but these identifications are theoretical diagnostics rather than validated assessments. Therefore, the accessibility-related findings in this study should be interpreted as preliminary indicators rather than comprehensive audits.

Fourth, generalization remains limited; although COP websites share structural characteristics, each COP edition adopts distinct editorial, political, and technological strategies, which may diminish the external validity of findings across conferences.

It is important to note that the GAI-based evaluation was conducted between March 10 and 15, 2025, whereas the human-based evaluation took place during COP30 in November 2025. Although the websites may have undergone incremental updates during this period, the study focused on comparing the overall usability patterns identified by GAIs and users rather than performing a page-by-page comparison of an identical website version. Therefore, the findings should be interpreted as complementary assessments of the platforms within their respective periods of analysis.

The GAI evaluations were based exclusively on the textual and structural representation of the websites (HTML/CSS metadata) and did not involve rendered visual inspection. Consequently, observations regarding visual hierarchy, aesthetics, or prominence should be interpreted as heuristic inferences rather than direct perceptual evaluations.

Although the recruitment strategy sought to maximize participant diversity, the exploratory sample was small and did not include representatives from all groups present at COP30, particularly Indigenous, quilombola, and riverside communities. Therefore, the findings should be interpreted as reflecting a range of participant perspectives rather than the full diversity of the conference audience.

Because the frequency analysis reflects the recurrence of analytically distinct findings rather than their practical impact, the interpretation of the results relied primarily on qualitative triangulation with the human evaluations.

6.2 Future works

The next steps of this study aim to advance in several directions. Longitudinal evaluations conducted before, during, and after major climate events would allow researchers to observe how usability perceptions evolve and how GAIs react to dynamic interface changes. Complementary analyses using automated accessibility checkers and screen-reader simulations could enrich the assessment of inclusive design, an essential dimension for global digital platforms. Additionally, examining how GAIs behave when interacting with multilingual, multimodal, or highly dynamic websites would help clarify their robustness in contexts where real-time updates, live streams, and diverse content formats are common.

Another promising line of inquiry involves developing hybrid workflows in which GAIs assist researchers by pre-categorizing issues, generating heuristic summaries, or supporting large-scale comparative analyses, enabling human evaluators to focus on more interpretive or contextual dimensions. Future studies should also explore how ongoing model updates affect reproducibility, particularly regarding subjective dimensions such as satisfaction, emotional clarity, and cultural appropriateness. Monitoring these evolutions will be key to determining whether GAIs can eventually support more stable and comprehensive evaluation frameworks for high-impact, globally oriented digital environments.

Declarations

Acknowledgements

The authors extend their gratitude to the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) for their financial and institutional support for this research. We also thank the Federal University of Pará (UFPA) and the Federal University of Santa Catarina (UFSC) for their institutional support. Special acknowledgment is given to the Government of Brazil and the Government of Pará, whose authorization and facilitation were essential for conducting the research within the COP30 spaces.

Funding

This research was funded by the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES).

Authors' Contributions

The contributions of the authors to this research were defined and described using the CRediT (Contributor Roles Taxonomy) framework (https://credit.niso.org/). The specific roles for this study were:

- **Conceptualization:** DTL developed the central research idea, defined the analytical scope involving both GAIs and human participants, and structured the comparative framework between COP29 and COP30 websites. MCRS and RCRP contributed to refining the conceptual foundation and the methodological alignment with HCI standards.
- **Methodology:** DTL designed the research protocol, including data collection procedures, prompt standardization, participant evaluation design, and the categorization framework based on ISO 9241-210. MCRS and RCRP provided methodological validation and guidance.
- **Data Collection:** DTL conducted all interactions with the selected GAIs and coordinated the usability evaluation sessions with human participants. AMM assisted in the data collection from the GAIs and the organization of the initial findings.
- **Data Curation:** DTL and AMM organized, cleaned, and systematized the qualitative dataset, structured the coding spreadsheet, and prepared the materials deposited in the Zenodo repository. AMM assisted in the initial compilation and verification of the data.
- **Formal Analysis:** DTL performed the directed content analysis, coded the GAIs' responses and participant feedback, and conducted the comparative interpretation. MCRS and RCRP revised the analytical coherence and contributed to identifying interpretive nuances.
- **Writing / Original Draft:** DTL drafted the full manuscript.
- **Writing / Review & Editing:** MCRS and RCRP critically revised the manuscript, provided substantive feedback, and contributed textual refinements.
- **Supervision:** MCRS provided primary supervision throughout the research. RCRP contributed to co-supervision, including theoretical grounding and methodological adjustments.

All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets (GAI outputs, user perceptions, and coded categories) generated and analysed during the current study are available in the public repository Zenodo repository: https://zenodo.org/records/21138198.

Further relevant information

This research involved the use of GAI tools during the empirical investigation stage. The qualitative data analysed were generated by the ChatGPT (OpenAI), DeepSeek, Gemini (Google), and Microsoft 365 Copilot platforms. Furthermore, the study adhered to ethical guidelines for research involving human participants by obtaining Free and Informed Consent from all adult volunteers and ensuring anonymity, as detailed in the Methods section.

Citation Diversity Statement: The authors of this paper acknowledge the existence of citation biases, which can lead to the underrepresentation of contributions from diverse research groups, particularly those from marginalized genders and non-dominant geographical contexts. As a research team aligned with the GranDIHC-BR agenda, we are committed to equity in scientific referencing. Our reference list includes foundational work primarily from North America and Europe, but also incorporates, contextually relevant contributions from researchers in Brazil and other non-dominant regions. Regarding gender diversity, an analysis of the first and last authors of the references revealed that 18 references feature men, 7 feature women, and 5 could not be categorized with certainty. We aim to continuously improve the equity of our reference lists in future works, actively seeking and prioritizing contributions from underrepresented groups in the field of Human-Computer Interaction.

References

Alsadi, A. and 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. DOI: https://doi.org/10.54489/ijtim.v3i1.195.

Araújo, M. F. B. d., Mota, M. P., and 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. DOI: https://doi.org/10.5753/ihc_estendido.2024.243962.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594. DOI: https://doi.org/10.1080/10447310802205776.

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

Bevan, N., Carter, J., and 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. DOI: https://doi.org/10.1007/978-3-319-20901-2_13.

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

Borges, J. M. and Araújo, R. D. (2024). Experiences and challenges of a redesign process with the support of an ai assistant on an educational platform. New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3702038.3702046.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101. DOI: https://doi.org/10.1191/1478088706qp063oa.

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: 22 August 2026.

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

Campos, T., Castello, M., Damasceno, E., and Valentim, N. (2025). An updated systematic mapping study on usability and user experience evaluation of touchable holographic solutions. *Journal on Interactive Systems*, 16(1):172–198. DOI: https://doi.org/10.5753/jis.2025.4694.

Carifio, J. and Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of social sciences*, 3(3):106–116. DOI: http://dx.doi.org/10.3844/jssp.2007.106.116.

Castells, M. (2008). The new public sphere: Global civil society, communication networks, and global governance. *The Annals of the American Academy of Political and Social Science*, 616(1):78–93. DOI: https://doi.org/10.1177/0002716207311877.

Cyr, D., Head, M., and Ivanov, A. (2006). Design aesthetics leading to m-loyalty in mobile commerce. *Information & management*, 43(8):950–963. DOI: https://doi.org/10.1016/j.im.2006.08.009.

de Carvalho, J. F. M., Moura, F. R. T., do Carmo Pereira, L. G., da Conceição Estevam, L., Negreiros, W. J. A. G., Alves, A. V. N., Pinto, L. V. L., dos Santos, A. J. S., Júnior, W. d. S. O., Cardoso, D. L., *et al*. (2026). Innovations in geometry teaching with miriti-board vr 2.0 glasses: Customizations and usability evaluation. *Journal on Interactive Systems*, 17(1):175–192. DOI: https://doi.org/10.5753/jis.2026.5755.

de Oliveira, L., Amaral, M., Bim, S., Valença, G., Almeida, L., Salgado, L., Gasparini, I., and da Silva, C. (2024). Grandihc-br 2025-2035 – gc3: Plurality and decoloniality in hci. In *Anais do XXIII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*, pages 1067–1085, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.1145/3702038.3702056.

Duarte, E. F., Toledo Palomino, P., Pontual Falcão, T., Porto, G. L. P. M. B., Portela, C. d. S., Ribeiro, D. F., Nascimento, A., Costa Aguiar, Y. P., Souza, M., Moutin Segoria Gasparotto, A., *et al*. (2024). Grandihc-br 2025-2035-gc6: Implications of artificial intelligence in hci: A discussion on paradigms ethics and diversity equity and inclusion. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*, pages 1–19. DOI: https://doi.org/10.1145/3702038.3702059.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3):379–383. DOI: https://doi.org/10.3758/BF03195514.

Fischer, M. and Lanquillon, C. (2024). Evaluation of generative ai-assisted software design and engineering: A user-

centered approach. In Degen, H. and Ntoa, S., editors, *Artificial Intelligence in HCI*, pages 31–47, Cham. Springer Nature Switzerland. DOI: https://dl.acm.org/doi/10.1007/978-3-031-60606-9_3.

Hassenzahl, M. and Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, 25(2):91–97. DOI: https://doi.org/10.1080/01449290500330331.

Hofstede, G. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Sage Publications, Thousand Oaks, 2 edition.

Hornbæk, K. and Hertzum, M. (2017). Technology acceptance and user experience: A review of the experiential component in hci. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):1–30. DOI: https://doi.org/10.1145/3127358.

International Organization for Standardization (2019). ISO 9241-210: Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. `https://www.iso.org/standard/77520.html`. Accessed: 31 August 2026.

Ivory, M. Y. and Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Comput. Surv.*, 33(4):470–516. DOI: https://doi.org/10.1145/503112.503114.

Kuric, E., Demcak, P., Krajcovic, M., and Lang, J. (2025). Systematic literature review of automation and artificial intelligence in usability issue detection. DOI: https://doi.org/10.48550/arXiv.2504.01415.

Lewis, J. R. (2018). The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590. DOI: https://doi.org/10.1080/10447318.2018.1455307.

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. DOI: https://doi.org/10.54097/8vpy7360.

Lima, D., Medeiros, A., de Cássia Paulino, R., and Seruffo, M. C. (2025). Can ai judge usability? a comparative analysis of generative tools on climate conference websites. In *Anais do XXIV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*, pages 497–517, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/ihc.2025.10805.

Lima, J. and de Lima, T. C. (2025). Blueprint para design participativo sensível ao contexto: Resposta aos desafios gc3 e gc4 do grandihc-br. In *Anais Estendidos do XXIV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*, pages 391–396, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/ihc_estendido.2025.16215.

Lincoln, Y. (1980). Guba. e.(1985). naturalistic inquiry. *Beverly Hills: Sage. LincolnNaturalistic Inquiry1985*. DOI: https://doi.org/10.1016/0147-1767(85)90062-8.

Liu, F. (2021). International usability testing: Why you need it. `https://www.nngroup.com/articles/why-international-usability-testing/`. Accessed: 23 August 2026.

Marcus, A. and Gould, E. W. (2000). Crosscurrents: cultural dimensions and global web user-interface design. *Interactions*, 7(4):32–46. DOI: https://doi.org/10.1145/345190.345238.

Meskó, B. and Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120. DOI: https://doi.org/10.1038/s41746-023-00873-0.

Miraz, M. H., Ali, M., and Excell, P. S. (2017). Multilingual website usability analysis based on an international user survey. *CoRR*, abs/1708.05085. DOI: https://doi.org/10.48550/arXiv.1708.05085.

Nielsen, J. (1994a). Estimating the number of subjects needed for a thinking aloud test. *International journal of human-computer studies*, 41(3):385–397. DOI: https://doi.org/10.1006/ijhc.1994.1065.

Nielsen, J. (1994b). Heuristic evaluation. In Nielsen, J. and Mack, R. L., editors, *Usability Inspection Methods*, pages 25–62. John Wiley & Sons, New York. DOI: https://dl.acm.org/doi/10.5555/189200.189209.

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

Nisbett, R. E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently... and Why*. The Free Press, New York.

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

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., *et al*. (2015). Promoting an open research culture. *Science*, 348(6242):1422–1425. DOI: https://doi.org/10.1126/science.aab2374.

Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, 16(1):1609406917733847. DOI: https://doi.org/10.1177/1609406917733847.

Oliveira, L. F. P. d. and 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).

Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in nursing & health*, 23(4):334–340. DOI: https://doi.org/10.1002/1098-240x(200008)23:4

Sauro, J. and Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2215–2224. DOI: https://doi.org/10.1145/1978942.1979266.

Sen, A. (2000). *Desenvolvimento como Liberdade*. Companhia das Letras, São Paulo.

Serra, L., Carvalho, L., Ferreira, L., Vaz, J., and Freire, A. (2015). Accessibility evaluation of e-government mobile applications in brazil. *Procedia Computer Science*, 67:348–357. DOI: https://doi.org/10.1016/j.procs.2015.09.279.

Shneiderman, B. and Plaisant, C. (2010). Designing the user interface: strategies for effective human-computer interaction. DOI: https://doi.org/10.1145/25065.950626.

Streiner, D., Norman, G. R., and Cairney, J. (2016). Health measurement scales: a practical guide to their development and use. *Aust NZJ Public Health*, 40(3):294–5. DOI: https://doi.org/10.1093/med/9780199685219.001.0001.

Tractinsky, N. (2018). The usability construct: a dead end? *Human–Computer Interaction*, 33(2):131–177. DOI: https://doi.org/10.1080/07370024.2017.1298038.

Vatrapu, R. and Pérez-Quiñones, M. A. (2004). Culture and international usability testing: The effects of culture in structured interviews. *CoRR*, cs.HC/0405045. DOI: https://doi.org/10.48550/arXiv.cs/0405045.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*. DOI: https://doi.org/10.48550/arXiv.2302.11382.