Insights and Implications of Evaluating Accessibility Compliance in AI-Generated Web Interfaces

Alexandra-Elena Guriţă
MintViz Lab, MANSiD Research Center
Ştefan cel Mare University of Suceava
Suceava, Romania
alexandra.gurita@student.usv.ro

Radu-Daniel Vatavu
MintViz Lab, MANSiD Research Center
Ştefan cel Mare University of Suceava
Suceava, Romania
radu.vatavu@usm.ro

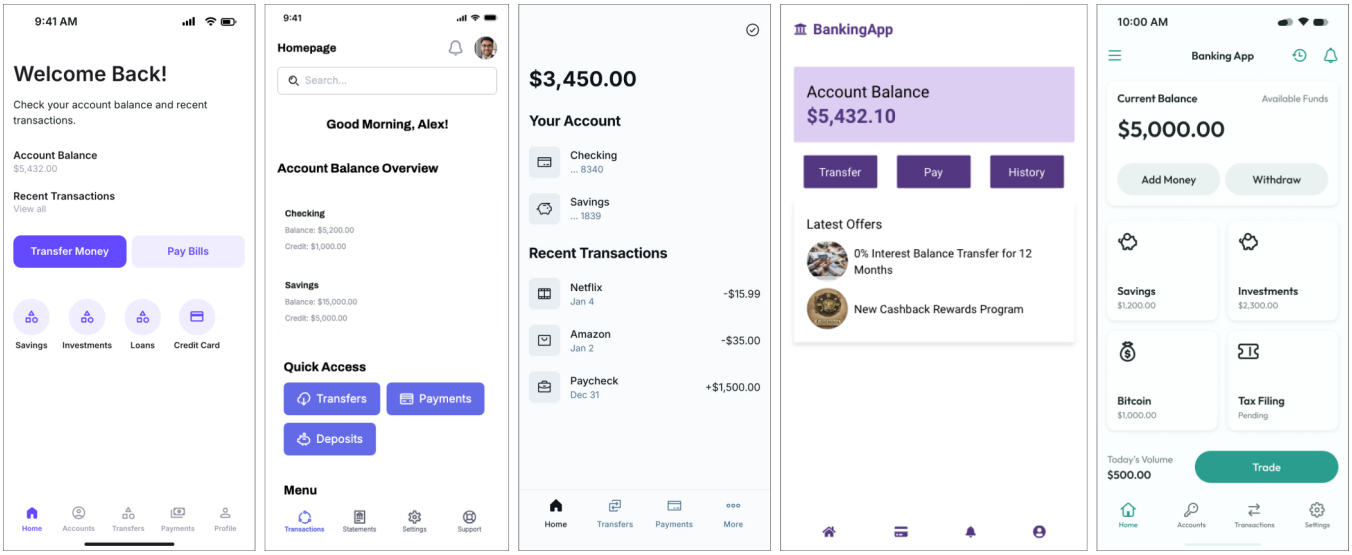

Figure 1: Examples of AI-generated user interfaces of a banking application across five different publicly available design tools examined in this work. Note the recurring design patterns featuring similar color schemes and navigation structures.

Abstract

The recent availability of AI-driven user interface generation tools necessitates a proper understanding of their capabilities to produce accessible designs in accordance to established standards. To this end, this paper reports results from an evaluation of fifty user interfaces, generated using five publicly available AI design tools, against WCAG 2.1 criteria (e.g., text contrast and target size), comparing both *accessibility-agnostic* and *accessibility-oriented* text prompts. Our analysis reveals moderate violation severity (M=0.47) on a scale from 0, none to 4, critical), with text contrast (M=1.08) and target size (M=0.86) emerging as primary yet fixable violations. Contrary to our expectations, the accessibility-oriented prompt did not improve compliance in our dataset, but actually reduced it (M=0.54 *vs.* M=0.39). However, three of the examined AI tools showed improved results through dialogue and iterative refinement.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW Companion '25, Sydney, NSW, Australia
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715552

CCS Concepts

• **Human-centered computing** → **Accessibility**; *User interface design*; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Accessibility; WCAG; AI-generated interfaces; User interface design; Generative AI; Design tools; Prompt engineering

ACM Reference Format:
Alexandra-Elena Guriţă and Radu-Daniel Vatavu. 2025. Insights and Implications of Evaluating Accessibility Compliance in AI-Generated Web Interfaces. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3715552

1 Introduction

The intersection of Artificial Intelligence (AI) and User Interface (UI) design has given rise to new tools that can automatically generate interfaces from text prompts [8,14]. Leading design platforms [16] are integrating such technological advances, while specialized tools [1,5,17,21] are implementing AI-first approaches to UI design. The emergence of AI-driven UI generation tools follows the broader trend of AI-generated content becoming increasingly prevalent in various domains, shaping the future of the web and the

Internet [6], where these tools serve as "autodesigners" [18]. However, this paradigmatic shift towards AI-generated UIs, achieved through natural language instructions, can make the design process become a series of brute-force trial and error [9,12], where designers—predominantly "visual thinkers" [13]—are constrained to iteratively refining text prompts to achieve desired outputs. Regrettably, this is a process that can feel random and unprincipled.

Although AI generation technology has evolved, statistics consistently reveal a tremendous accessibility gap, e.g., the WebAIM Million 2024 report [20] indicates that 95.9% of websites fail to meet WCAG standards. In this context, the emergence of AI-generated UIs introduces new questions about accessibility compliance. Unlike traditional accessibility implementation, which depends on practitioners' knowledge and conventional design tools, it remains unclear how prompt engineering influences the accessibility of AI-generated UIs to arrive at WCAG-compliant outcomes. To address this gap, we present an evaluation of UIs generated with five AI design tools, comparing their WCAG 2.1 compliance, at levels A, AA, and AAA, with both *accessibility-agnostic* and *accessibility-oriented* text prompts. Our specific research questions are:

- To what extent do AI-generated UIs comply with current WCAG accessibility standards?
- How does explicit inclusion of accessibility requirements in prompts, i.e., *accessibility-oriented prompt engineering*, increase the WCAG compliance of AI-generated UIs?
- What accessibility patterns emerge across AI-generated UIs?

We believe answering these questions has important implications for tool developers, designers, and researchers working at the intersection of generative AI and UI design.

2 Study

We conducted a study to evaluate the accessibility compliance of UIs generated by various AI-driven design tools.

2.1 User Interface Generation

2.1.1 Phase 1: Prompt engineering. We designed two prompts: (1) an *accessibility-agnostic* prompt representing a baseline instruction requesting a banking application homepage design without any explicit reference to accessibility specifications, serving as the control condition in our evaluation, and (2) an *accessibility-oriented* prompt, an enhanced version of (1), incorporating five WCAG criteria—1.4.1A, 1.4.11AA, 1.4.12AA, 1.4.3AA, 2.5.5AAA, covering color use, contrast ratios, text spacing, and target sizes—which we selected based on their testability in static image outputs. To ensure prompt consistency while respecting tool limitations, we maintained a 300-character limit and used natural language adaptations of WCAG specifications. The character limit was imposed by the maximum input length supported by one of the tools included in our study, a constraint that we applied across all tools to maintain evaluation consistency. The following text prompts were used across all design tools utilized in this study:

> **(1) Accessibility-agnostic prompt:**
> "Design the homepage of a banking app."
>
> **(2) Accessibility-oriented prompt:**
> "Design the homepage of banking app with 4.5:1 text contrast (3:1 for 18pt+) AND color isn't the sole info indicator (uses icons, patterns, or text). Clickable elements are 44x44 pixels, UI components have 3:1 contrast. Text spacing is 1.5x line height, 2x paragraph."

2.1.2 Phase 2: User interface generation. We implemented a factorial design involving two independent variables, Prompt (nominal variable with two conditions, presented above) and Tool (nominal variable with five conditions, represented by five design platforms—Figma, GalileoAI, Uizard, Banani, and Visily[1]—selected based on their public availability and capability to generate complete UI designs from text prompts at the time of our study); see Figure 2 for several examples of UIs in our dataset that were generated using *accessibility-agnostic* (top) and *accessibility-oriented* (bottom) text prompts. The Tool variable represents only a means of generating UIs for our evaluation from different sources, and we are not interested in its specific conditions. Thus, in our analysis, we treat it as a random effect and aggregate its values. Each tool was used to generate ten different UIs, resulting in a dataset of of size fifty, all of which we verified to be complete and valid for evaluation.

2.2 Evaluation

2.2.1 Scoring and dependent variables. We implemented a systematic accessibility evaluation method, building on prior work [2, 11], through a 5-point severity scale, representing our Violation-Severity measure. We used this scale to assess the WCAG compliance of the UIs in our dataset from 0 (no violation) to 1 (minor, cosmetic issues), 2 (moderate, partial accessibility barrier), 3 (serious, significant barrier), and 4 (critical, complete barrier). This granular scoring enabled precise measurement of accessibility compliance variations, particularly valuable when analyzing UIs containing multiple instances of similar elements (e.g., buttons, text blocks) with inconsistent accessibility implementations. As shown in Figure 1, contrast ratios can vary within the same interface. For example, the second UI in Figure 1 generated primary buttons at 4.41:1 and navigation at 6.53:1 contrast ratio, while the rightmost UI has the "Add Money" and "Withdraw" buttons at 18.43:1 but the navigation and "Trade" buttons at 3.32:1, failing WCAG 1.4.3 AA.

Using the severity scores, we calculated a derived measure, expressed as a percentage, where we considered an interface to be in violation if it received any severity score greater than 0:

$$\text{Violation-Rate} = [\text{Violation-Severity} > 0] \cdot 100\% \quad (1)$$

where $[\cdot]$ represents a logical mask that evaluates to 1 if the inner expression is true and 0 otherwise. Both of our previous examples would result in a Violation-Rate of 100%, making this measure more conservative than the granular, 0 to 4, Violation-Severity.

2.2.2 Evaluation process. Two independent accessibility design experts (holding 6 and 7 years of experience, respectively) conducted the evaluations using a standardized protocol, based on documented methodologies [3,4,15], integrating our specific evaluation measures, as follows. To minimize bias, the evaluators were not informed about which tool had generated which UI. They assessed each interface against the five selected WCAG criteria, documenting specific violations and scoring their severity levels from 0 to 4.

[1] https://www.figma.com, https://www.usegalileo.ai/explore, https://uizard.io, https://www.banani.co, https://www.visily.ai

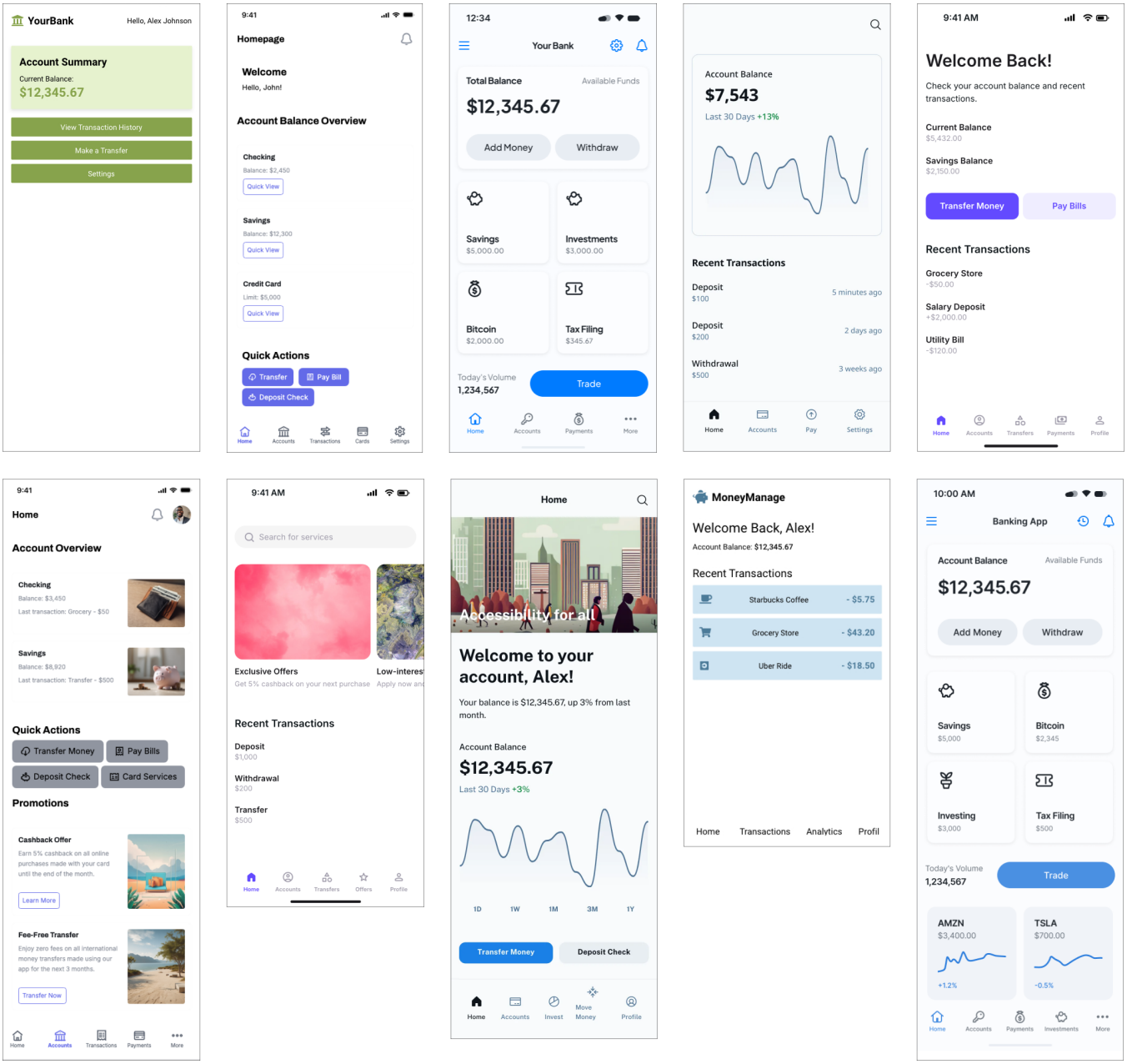

Figure 2: Screenshots of selected UIs in our dataset, generated using *accessibility-agnostic* (top) and *accessibility-oriented* (bottom) text prompts; see Figure 3 and Table 1 for accessibility compliance evaluation results of these two PROMPT type conditions in terms of violation rates and severity levels with respect to WCAG criteria.

For objective measurements, we used the Stark plugin[2] in Figma to evaluate text and non-text contrast ratios for UI elements, while other criteria required expert visual inspection [19]. Each evaluator rated the severity of accessibility violations in all UIs in our dataset, leading to a total of 2 (evaluators) × 2 (prompt types) × 5 (AI design tools) × 10 (repetitions) = 200 scores.

[2]https://www.figma.com/community/plugin/732603254453395948/stark-contrast-accessibility-checker

3 Results

3.1 Violation Rate

Our evaluation of fifty AI-generated UIs (see Figure 1 for examples in our dataset) revealed varying accessibility compliance, averaged across all tested tools. UIs generated using the control, *accessibility-agnostic* text prompt exhibited an average violation rate of 21% across all five WCAG criteria. Text contrast violations were the

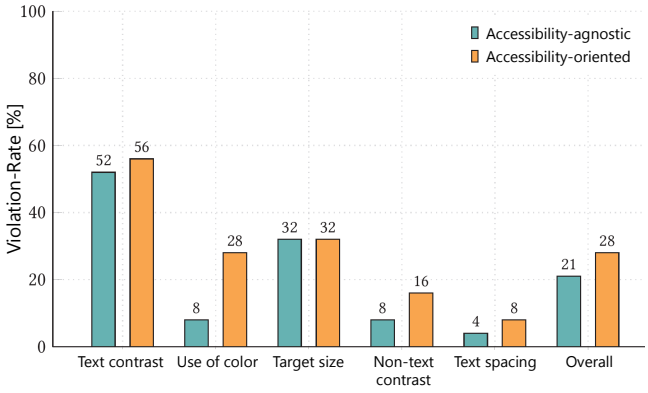

Figure 3: Violation rates according to various WCAG criteria and Prompt types; see Table 1 for severity scores.

Table 1: Average violation severity scores according to Prompt type; see Figure 3 for violation rates. *Note:* lower scores denote better performance; the minimum score is 0 (no violation) and the maximum is 4 (critical barrier).

Prompt type	TC†	UC†	TS†	NTC†	TS†	Mean
Accessibility-agnostic	0.92	0.08	0.80	0.08	0.08	**0.39**
Accessibility-oriented	1.24	0.28	0.92	0.16	0.08	**0.54**
Mean	**1.08**	**0.18**	**0.86**	**0.12**	**0.08**	**0.47**

WCAG criteria: †TC–text contrast, UC–use of color, TS–target size, NTC–non-text contrast, and TS–text spacing.

most prevalent (52% of UIs), followed by target size (32%), use of color (8%) and non-text contrast (8%), while text spacing showed the lowest violation rate at 4%. The *accessibility-oriented* prompt led to a slightly higher violation rate of 28%, caused primarily by an increase in use of color violations, from 8% to 28%; see Figure 3.

3.2 Violation Severity

Text contrast violations showed the highest average severity score (M=1.08), particularly with the *accessibility-oriented* text prompt (M=1.24) compared to the *accessibility-agnostic* one (M=0.92); see Table 1. Target size was the second most severe issue (M=0.86), whereas use of color, non-text contrast, and text spacing showed much lower severity scores (M=0.18, 0.12, and 0.08 respectively), although the *accessibility-oriented* prompt consistently exhibited slightly higher severity scores in these criteria. Overall, the severity level across all five criteria was low—M=0.47 on the 0 to 4 scale—but the frequency of failure, assessed with the conservative violation rate measure, presented in Figure 3, was high.

4 Discussion

We use our empirical findings to propose practical implications for AI-driven design tools used for UI generation from the perspective of the accessibility compliance of the interfaces they produce.

4.1 The Current State of AI-Generated User Interface Accessibility

Our findings revealed persistent accessibility challenges in AI-generated UIs, with varying severity scores across different WCAG criteria. Among these, text contrast emerged as the most problematic issue, with the highest average severity score. This finding suggests that current AI tools, despite their capability of generating visually appealing interfaces, may not effectively encode fundamental accessibility principles about the visual contrast and size of the UIs' constituting elements. This aligns with broader observations about AI-generated content [10], where AI models often prioritize aesthetic patterns over functional requirements. The distribution of violations across different WCAG criteria provides insights into specific challenges to address in future work: while some technical requirements like text spacing showed lower severity scores, criteria such as target size proved more challenging for the AI design tools evaluated in our study to implement correctly.

4.2 Prompts with Explicit Accessibility Requirements Are Not Much Benefit

Contrary to our expectation, UIs generated with an *accessibility-oriented* text prompt showed a slightly higher average violation rate compared to the control, *accessibility-agnostic* condition. These findings suggest that just by including accessibility requirements in text prompts may not be sufficient to improve WCAG compliance of the resulting UIs and, for some criteria such use of color and text contrast, might even lead to trade-offs in accessibility implementation. We believe that this result may be caused by current AI models struggling to simultaneously optimize for multiple accessibility criteria explicitly integrated into the prompt, potentially due to their specific training settings—where they are capable of reproducing accessible patterns, but incapable of thoughtful deviation when specific contexts might demand it [7] —or their limited capability of processing diverse requirements within a single text prompt.

4.3 Inconsistent Application of Accessibility Criteria in AI-generated User Interfaces

We identified a notable distinct failure pattern across multiple tools, pointing to a specific area where AI models may struggle with accessibility requirements—inconsistent application of accessibility criteria over the same type of UI element in multiple places in the generated interface. We observed this inconsistency in the generated elements of the UIs in our dataset manifesting as (i) varying contrast ratios or touch target sizes for similar UI elements, visible in the second and rightmost interfaces shown in Figure 1, and (ii) different approaches to redundant indicators for color-based information, as shown in the fourth interface in Figure 1.

4.4 Solving Accessibility Issues through Iterative Prompting

Our analysis revealed that three out of the five tools evaluated in our study supported accessibility improvements through iterative dialogue. While initial responses from these tools typically indicated an inability to address accessibility issues, continued designer engagement led to successful resolution. The median resolution

occurred at the third interaction attempt. For example, one tool improved its violation severity score from 0.84 to 0.32 (lower is better, see Subsection 2.2 for details on this score) through three iterations, where we achieved better results by requesting specific adjustments (e.g., "Please increase the contrast ratio of the navigation buttons") rather than using general accessibility improvements. This finding suggests a key limitation in current text prompt engineering approaches: while accessibility knowledge appears to be encoded in the AI models to some extent, it requires specific elicitation through persistent follow-up interaction. This aligns with prior work on interactive prompt refinement [9], extending to accessibility-specific contexts. However, this finding also raises concerns about the cognitive load placed on designers and developers, which the scientific literature has already highlighted as a challenge [13]. Rather than relying solely on initial prompts, practitioners should engage in systematic iteration and refinement of the UIs during the generation process, where human expertise and oversight remain essential.

5 Conclusion and Future Work

Our evaluation of AI-generated UIs revealed a spectrum of accessibility compliance, with particular challenges in text contrast and target size requirements. Surprisingly, explicitly incorporating accessibility requirements in prompts did not consistently improve compliance and even led to increased violation rates. While most tools could resolve accessibility issues through iterative dialogue, this suggests both an opportunity and a challenge in effective implementation of accessibility knowledge within AI design tools. These findings should be considered within the context represented by limitations of our study, where we focused on static interface elements and excluded dynamic accessibility features, evaluated only visually assessable WCAG criteria, and the tools' default color combinations may have influenced the accessibility outcomes.

Our results open several directions for future work, as follows: (1) developing specialized accessibility-aware prompting frameworks that better handle multiple WCAG criteria simultaneously, (2) studying how practitioners modify AI-generated UIs to meet accessibility compliance through iterative refinement, (3) incorporating automated evaluation features for real-time accessibility feedback, and (4) investigating methods to improve AI models' understanding of accessibility requirements through specialized training. We also recommend future work that replicates and consolidates our findings by involving larger datasets and more varied accessibility-oriented text prompts, such as prompts designed to generate UIs for users with specific sensory or motor abilities.

Acknowledgments

This work was supported by a grant of the Ministry of Education and Research, CCCDI-UEFISCDI, project number PN-IV-P7-7.1-PTE-2024-0434, within PNCDI IV.

References

[1] Banani. 2024. UI Design AI Generator. https://www.banani.co/product/ai-ui-generator

[2] Thiago Jabur Bittar, Leandro Agostini do Amaral, and Renata Pontin de Mattos Fortes. 2011. AccessibilityUtil: A Tool for Sharing Experiences about Accessibility of Web Artifacts. In *Proceedings of the 29th ACM International Conference on Design of Communication (SIGDOC '11)*. ACM, New York, NY, USA, 17–24. doi:10.1145/2038476.2038480

[3] Giorgio Brajnik. 2008. Beyond Conformance: The Role of Accessibility Evaluation Methods. In *Proceedings of the Web Information Systems Engineering – WISE 2008 Workshops*, Sven Hartmann, Xiaofang Zhou, and Markus Kirchberg (Eds.). Springer, Berlin, Heidelberg, 63–80. doi:10.1007/978-3-540-85200-1_9

[4] Tânia Frazão and Carlos Duarte. 2020. Comparing Accessibility Evaluation Plugins. In *Proceedings of the 17th International Web for All Conference (W4A '20)*. ACM, New York, NY, USA, Article 20, 11 pages. doi:10.1145/3371300.3383346

[5] Galileo. 2024. Galileo AI's Public Beta & Seed Round Led by Khosla Ventures. https://www.usegalileo.ai/blog/seed

[6] Wensheng Gan, Zhenqiang Ye, Shicheng Wan, and Philip S. Yu. 2023. Web 3.0: The Future of Internet. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*. ACM, New York, NY, USA, 1266–1275. doi:10.1145/3543873.3587583

[7] Alexandra-Elena Guriță and Radu-Daniel Vatavu. 2025. Breaking Bad (Design): Challenging AI User Interface Accessibility Guardrails. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA'25)*. ACM, New York, NY, USA, 7 pages.

[8] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. ACM, New York, NY, USA, Article 35, 8 pages. doi:10.1145/3491101.3503564

[9] Vivian Liu and Lydia B. Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, Article 384, 23 pages. doi:10.1145/3491102.3501825

[10] Jakob Nielsen. 2023. AI: First New UI Paradigm in 60 Years. https://www.nngroup.com/articles/ai-paradigm

[11] George Moreno De Oliveira and Ingrid Teixeira Monteiro. 2023. Development and Evaluation of the Plugin for Figma for Accessibility Documentation for Interfaces - DAI. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems (IHC '23)*. ACM, New York, NY, USA, Article 37, 11 pages. doi:10.1145/3638067.3638102

[12] Jonas Oppenlaender. 2023. A Taxonomy of Prompt Modifiers for Text-to-Image Generation. *Behaviour & Information Technology* 43, 15 (2023), 3763–3776. doi:10.1080/0144929x.2023.2286532

[13] Hyerim Park, Joscha Eirich, Andre Luckow, and Michael Sedlmair. 2024. "We Are Visual Thinkers, Not Verbal Thinkers!": A Thematic Analysis of How Professional Designers Use Generative AI Image Generation Tools. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction (NordiCHI '24)*. ACM, New York, NY, USA, Article 35, 14 pages. doi:10.1145/3679318.3685370

[14] Savvas Petridis, Michael Terry, and Carrie J Cai. 2024. PromptInfuser: How Tightly Coupling AI and UI Design Impacts Designers' Workflows. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. ACM, New York, NY, USA, 743–756. doi:10.1145/3643834.3661613

[15] Helen Petrie and Nigel Bevan. 2009. The Evaluation of Accessibility, Usability, and User Experience. In *The Universal Access Handbook* (1st ed.), C. Stephanidis (Ed.). CRC Press, Boca Raton. http://www.crcpress.com/product/isbn/9780805862805

[16] Kris Rasmussen. 2024. Meet Figma AI: Empowering Designers with Intelligent Tools. https://www.figma.com/blog/introducing-figma-ai

[17] Uizard. 2021. Uizard Launches World's First AI-Powered Design Tool for Non-Designers. https://uizard.io/blog/uizard-launches-worlds-first-ai-powered-design-assistant

[18] Uizard. 2024. Designing with Text Prompts: Create UI Designs Faster than Ever Before. https://uizard.io/blog/create-ui-designs-using-text-prompts

[19] W3C Web Accessibility Initiative. 2024. Using Combined Expertise to Evaluate Web Accessibility. https://www.w3.org/WAI/test-evaluate/combined-expertise

[20] WebAIM. 2024. The WebAIM Million: The 2024 Report on the Accessibility of the Top 1,000,000 Home Pages. https://webaim.org/projects/million

[21] Jordan Woods. 2024. Announcing the Launch of Visily Pro. https://www.visily.ai/blog/announcing-the-launch-of-visily-pro