Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Analysis of LLM-Based Conversational Agents as an Assistive Technology for Web Interaction

Giuseppe Della Penna
University of L'Aquila
Marina Buzzi
marina.buzzi@cnr.it
National Research Council
Barbara Leporini
University of Pisa

Research Article

Keywords: Generative AI, Web Accessibility, Assistive Technology, LLM-based Conversational Agents

Posted Date: June 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9581264/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Comparative Analysis of LLM-Based Conversational Agents as an Assistive Technology for Web Interaction

Giuseppe Della Penna[1†], Marina Buzzi[2*†], Barbara Leporini[2,3†]

[1]Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, via Vetoio, Coppito, L'Aquila, 67100, Italy.
[2]CNR-IIT, via Moruzzi 1, Pisa, 56124, Italy.
[3]Department of Computer Science, University of Pisa, via Fibonacci, Pisa, 56124, Italy.

*Corresponding author(s). E-mail(s): marina.buzzi@cnr.it;
Contributing authors: giuseppe.dellapenna@univaq.it; barbara.leporini@unipi.it;
†These authors contributed equally to this work.

Abstract

This study investigates whether LLM-based Conversational Agents (LLMCAs) could be employed as assistive technologies, helping anyone navigate a website by explaining its content and interactive features, and suggesting actions to accomplish specific goals. This would make the web more accessible especially to seniors, blind users and people with mild cognitive impairment. Specifically, the study examines whether such tools can analyze the source code of a webpage to identify its structure and interactive elements, understand which interactive components are required to support users in achieving their goals, and act effectively as assistive technologies. In addition, the work compares leading tools to assess differences in their effectiveness and behavior in supporting accessible web interaction.

The work presents two empirical studies in which four popular LLMCA tools are evaluated on a set of public websites, in order to assess their ability to understand web page structures and interactive components and assist users in their tasks. The evaluation combines qualitative and quantitative measures, focusing on accuracy, completeness, and practical usefulness of the responses.

Results show a good degree of pertinent and good-quality answers, even on websites with a poorly accessible structure. The user-friendly way to interact with a LLMCA makes it much comfortable to use for the average user. However, the study also highlights that some LLMCAs appear more 'skilled' in accessibility tasks than others.

LLMCAs show promising potential not only to complement existing assistive technologies but also to support or replace the human assistance that is often required for many users to effectively access the web.

Keywords: Generative AI, Web Accessibility, Assistive Technology, LLM-based Conversational Agents

1 Introduction

Nowadays websites are an essential component of our lives. Several daily activities, including study, work, leisure, and social activities, as well as eGovernment services necessary for several tasks (tax payment, eHealth services, etc.), are managed online. Unfortunately, these services are not

accessible and easy to use for everyone, especially seniors and people with mild cognitive impairment. Although the first version of the WCAGs (Web Content Accessibility Guidelines) was published more than 25 years ago and accessibility requirements have since been reinforced by national and international legislation, many public service websites still reach only partial compliance, negatively impacting people with disabilities and seniors [10, 21]. Factors such as complexity of interaction, feelings of technical inadequacy, authentication process complexity, and strict time requirements can represent obstacles for seniors [5].

Moreover, web-based applications, as well as other digital platforms, are becoming increasingly complex, incorporating a growing number of multimedia elements and interactive components. To accomplish tasks such as scheduling a medical appointment or gathering information about benefits for first-time homebuyers, users often need to perform numerous actions, including navigating through several links, browsing multi-layered pages, and filling out lengthy and complex forms. As a result, this level of interaction can be tiring and frustrating for users with basic digital skills, for seniors, or individuals who rely on assistive technologies. Moreover, even for not skilled screen reader users, navigating content sequentially can be time-consuming, requiring more effort to use specific commands and increasing cognitive overload in the interaction [39].

LLM-based conversational agents (LLMCAs) can play a key role in bridging users and web interfaces by simplifying user interaction. Current assistive technologies in this field typically guide the user through the website structure elements using voice or visual hints. This kind of technology is essentially targeted at visually impaired subjects. However, web accessibility is a wider issue that also includes people with cognitive impairment and, more generally, people that are not "skilled" with modern web applications, like seniors.

We want to investigate if a LLMCA can guide the user through a site, describe the general site contents, explain the site interface and interactive elements, and suggest actions or procedures to accomplish a goal. These assistive capabilities clearly target visually impaired people, but also a wider audience. Many current LLMCA-driven tools simply look at a web interface as an image or as simple textual content (i.e., without structure). On the other hand, the LLMCA could enhance its interpretation by "reasoning" on the *structure*, *content* and *appearance* of web pages, thus effectively being able to empower humans (and in particular seniors and people with disabilities) when interacting with the web technology. However, to achieve this goal, these tools need to access and analyze the complete page *source code* (i.e., at least HTML and possibly CSS) to identify interactive elements and understand how they may contribute to the achievement of a task.

In this work, we present the results of two empirical studies conducted by querying four popular LLMCAs on a set of 13 websites that offer services and useful information for citizens. The aim is to assess the ability of LLMCAs to *understand* the websites' interface structure and interactive elements and, therefore, effectively guide users to accomplish their tasks. Specifically, the research questions that have driven our study can be summarized as follows:

RQ1: Are AI-powered tools capable of analyzing the source code of a web page to recognize its structure and interactive elements?

RQ2: Can AI-powered tools understand the interactive elements necessary to support users achieving their goals?

RQ3: Do leading AI-powered tools differ in their ability to analyze web page source code and understand its interactive elements?

If research questions RQ1 and RQ2 can be answered positively, even partially, it would suggest that AI-powered tools, or some of them depending on the answer to RQ3, have the potential to assist users in web navigation. This could make them valuable assistive technologies to support people in browsing more efficiently and with greater ease.

At the time of our first experiments, starting in 2024, the only available way to feed the LLMCAs with a web page source code was to manually paste it in the conversation as a prompt. However, manually providing such a source as part of the user prompt would require too much effort for any user and a significant cost in terms of *tokens* passed to the LLMCA. We feel this aspect is particularly important, since an assistive technology should not charge the user with additional costs or cognitive loads. Therefore, we initially focused

our empirical study for RQ1 and RQ2 on *Copilot*, as it was the only one that, integrated as an extension into browsers like Edge and Chrome, was able to directly read the code of the current web page using it as the "query context", providing easier access to users with no additional costs. The results of these first experiments (presented in Della Penna et al. [8]) have been very promising: as shown in Section 3, Copilot showed good abilities as an advanced assistive tool. However, these results were only preliminary, since we felt that other LLMCAs like *ChatGPT* or *Gemini* were quickly evolving.

The turning point in our study has been the introduction and wide availability of the Retrieval-Augmented Generation (RAG) model [17, 24]. Indeed, even if modern LLMCAs cannot directly access the web or update their internal parameters during inference using the web pages source code, they now achieve live web integration through an external *orchestration* pipeline. This pipeline can retrieve and inject external text like web page content into the model's *input context*, e.g., augmenting the prompt and giving the external content to the LLMCA as part of it. In other words, external knowledge can now be *automatically* used to enhance the model's input of *all* the LLMCAs, similarly to what we were able to do with Copilot in our first experiments. Therefore, we repeated an updated set of experiments at the end of 2025, this time comparing the results deriving from the same questions asked to four different leading LLMCAs. The new results confirmed that a LLMCA can be a valuable assistive tool, and showed the potential and limitations of the analyzed tools when applied to well-known websites.

The paper is organized in five sections. Section 2 lists some of the most interesting works in the field of assistive technologies and applications of AI tools to accessibility tasks. Sections 3 and 4 report our two empirical studies on the application of LLMCAs to web accessibility, whose results are then briefly discussed in Section 5. Finally, Section 6 outlines our conclusions and future work in this field.

2 Related work

Research in digital accessibility increasingly recognizes that persistent inequalities in access to online services are caused not only by usability issues but also by broader structural and socio-economic factors, such as digital literacy gaps, affordability constraints, accessibility barriers, and infrastructural limitations. A recent study highlights how these multidimensional determinants continue to shape unequal access to digital ecosystems among vulnerable populations, including seniors and persons with disabilities [36]. Despite the rise of digital public and private services, seniors and individuals with functional impairments still experience disproportionate difficulties due to complex interfaces and insufficiently inclusive design approaches [11]. Within this scenario, accessibility, usability, and user experience (UX) remain foundational for enabling equitable interaction with web services.

A growing portion of literature explores the potential of generative AI (GenAI) and Large Language Models (LLMs) to enhance accessibility. For instance, Patel et al. [33] propose an LLM-based system for domain-general image captioning, achieving high accuracy in generating descriptions useful to visually impaired users. López-Gil and Pereira [26] investigate the automation of web accessibility evaluations using LLMs to address WCAG success criteria traditionally assessed manually, thereby reducing time and costs for practitioners. Ghosh et al. [12] leverage LLMs to enhance screen-reader interaction by interpreting and performing actions on behalf of users, which reduces cognitive load and mitigates inconsistencies between interfaces. In this way, the system enables consistent access across different applications and reduces obstacles of diverse user interfaces as well as the burden to learn new layouts and structures, improving user efficiency in performing tasks. These efforts collectively underscore the emerging role of generative systems in lowering interaction barriers and ensuring more consistent accessibility across platforms.

Beyond AI-driven augmentation, research has long explored methods to simplify web interaction for older adults. Prior work includes design guidelines tailored to seniors' cognitive and perceptual needs [28, 34], as well as web simplification

approaches that restructure pages to reduce complexity and enhance interpretability [19, 30]. Web simplification typically removes unnecessary components, streamlines navigation structures, and employs simple language, improving comprehension for seniors and individuals with mild cognitive impairment. Tools supporting text simplification have also been evaluated with older adults and people with intellectual disabilities, showing positive user experiences [30].

Research on easy-to-read and text simplification strategies confirms that linguistic and structural simplification is a key enabler of digital inclusion, especially for seniors and users with cognitive limitations, providing a strong conceptual foundation for recent LLM-based approaches to automatic content simplification and personalization [13]. For instance, Uricchio et al. [41] investigates the use of LLMs for automatic sentence simplification based on easy-to-read principles [40], showing promise for generating accessible cultural heritage descriptions.

AI has been recognized as a valuable enabler of active participation in the digital society. Wang [42] argues that AI systems, especially generative models, hold substantial potential to support older adults and people with disabilities across a variety of tasks, although their full capabilities remain underexplored. Domain-tailored generative systems, such as those discussed by Jiang et al. [18], highlight the growing interest in customizing LLMs for more accurate and context-sensitive support. Recent research further confirms the growing role of AI in reducing cognitive load and improving trust and autonomy for older adults. Explainable AI significantly enhances perceived usability and trust among elderly users, directly addressing barriers related to cognitive overload and acceptance of intelligent systems [25].

Research targeting seniors' autonomous navigation of the web is still sparse, despite the rapid demographic growth of ageing populations and the associated need for tools that sustain healthy, independent, and socially connected lifestyles. Recent studies examine AI's broader role in ageing. Yau and Shen [44] show that older adults' acceptance of AI-driven health technologies depends on factors such as trust, perceived usefulness, clarity of purpose, perceived autonomy, and concerns related to transparency and control. More, Schoenborn et al. [37] highlight significant misalignment between priorities expressed by older adults, caregivers, clinicians, and developers, recommending inclusive and participatory approaches in the design of AI for ageing. Seniors often express ambivalent attitudes marked by interest, uncertainty, and concerns over privacy, regulation, and impacts on human connection [44].

These findings suggest that individual attitudes toward AI adoption among older adults are not shaped only by usability or perceived usefulness but are also influenced by broader systemic conditions, including how AI technologies are designed, regulated, and governed.

At a policy level, Chu et al. [6] remark that biased datasets, limited representation of older adults in AI development, and low technical literacy increase digital exclusion, thus reinforcing the need for inclusive data practices and accessible AI governance.

The integration of generative and autonomous agents into web accessibility represents a paradigm shift. Traditionally, web accessibility for seniors relies on static tools (such as screen readers or accessible templates accommodating size and contrast needs) that require the user to adapt to the interface. Today, AI agents are flipping this dynamic, allowing the interface to adapt to the user in real-time. Agents are transforming the web experience for older adults in different areas.

In the domain of web accessibility, generative agents are evolving from mere assistive overlays into autonomous intermediaries. By dynamically restructuring web content to meet the cognitive and physical needs of seniors, these agents facilitate a smooth digital experience [20].

Concerning personalization and content simplification, one of the most significant barriers for seniors is cognitive overload, dense layouts, technical jargon, or complex navigation. GenAI can perform real-time simplification, multimodal transformation, and visual aid creation. Generative agents can intercept complex web text and convert it instantly into simpler language or bullet lists, simplifying the reading level without losing the core meaning [31].

The effectiveness of agents in transforming complex websites into accessible formats is supported by recent work on generative UI architectures [3]. These systems are capable of dynamically manipulating the user interface to generate

dynamic tooltips and semantic simplifications. Such interventions significantly reduce cognitive load, addressing a critical necessity highlighted in disability and digital ageing research in relation to accessibility guidelines [4]. Agents can autonomously convert a text-heavy website into an audio summary or a simplified visual flow, adapting content to seniors with visual or cognitive decline or can create visual aids: if a button or icon is ambiguous, a generative agent can provide a descriptive tooltip or even redesign the element on-the-fly to be more recognizable to the specific user.

Autonomous navigation supported by conversational interfaces is another key aspect. For seniors with motor impairments or having difficulties with the point-and-click navigation devices, autonomous agents act as a bridge, enabling task-oriented browsing and predictive assistance. For example, instead of navigating through numerous sub-menus to pay a utility bill, a senior could tell an autonomous agent to pay with his saved card. The agent then autonomously executes the sequence of clicks and form-fills across the web interface. Autonomous agents also learn a user's recurrent difficulties. If a user consistently has problems with a specific checkout flow, the agent can anticipate the need and offer guidance or highlight the next required step.

Last, agents contribute to bridging the digital literacy gap. Seniors often experience anxiety when a website changes (due to updates to UI/UX). Generative agents can mitigate this emotional factor by providing a layer of consistency in different ways. Through adaptive UI, the agent can skin different websites into a consistent, familiar interface that the senior has already mastered. This reduces the learning curve for every new site visited. This could be especially valuable for the blind who have to memorize UI mentally. The user interacts with agents and asks for support in natural language. The agent can highlight the UI element or explain the change.

Mitra [29] explores the potential of generative AI and the metaverse to create hybrid physical-digital environments that support new forms of interaction and assistance. Zhou et al. [45] argue that natural-language-driven autonomous agents could significantly empower individuals by executing tasks efficiently and reducing interaction issues. Yet their effectiveness remains difficult to evaluate: Deng et al. [9] show that many agents operate under unrealistic static conditions, while Shridhar et al. [38] and Yao et al. [43] highlight the limitations of oversimplified environments that fail to capture real-world task diversity. To address this gap, Hambley et al. [15] introduced a reproducible environment for evaluating web-based language agents. Their findings indicate that complex tasks remain challenging for current models, with end-to-end success rates around 14.5%, far below human performance. However, in the last two years, LLMs have progressed very fast.

Conversational agents have gained prominence as accessible, natural interfaces for interacting with digital services. Pucci et al. [35] observe that, despite promising use cases, limited research has explored how conversational interfaces can make the web truly accessible for all users. A recent systematic review by Huq et al. [16] analyzes conversational and agent-based systems designed for seniors and cognitively impaired users, identifying essential features and outlining current shortcomings. An et al. [1] further argue that promoting lifelong learning and interdependent social support can help overcome older adults' resistance to conversational agent adoption. They emphasize personalization, such as tailoring multimodal interactions to user preferences, as a key factor for improving acceptance.

Recent developments in conversational web navigation take this further by enabling digital agents to control browsers and complete tasks through dialogue. Lù et al. [27], for example, introduce WEBLINX, a large-scale benchmark showing the need for multimodal models capable of generalizing across novel web contexts.

Research on the use of Large Language Models (LLMs) for enhancing web accessibility has expanded in recent years, though existing work primarily focuses on automated accessibility evaluation or remediation rather than conversational assistance for real-world web navigation. Several studies explore how LLMs can support or augment traditional WCAG compliance checks. For example, López-Gil and Pereira [26] investigate whether LLM-based systems can automatically evaluate WCAG success criteria that typically require manual inspection, showing that LLMs can successfully detect accessibility issues overlooked by automated tools, achieving an overall detection rate of 87.18% across selected criteria.

While this work highlights the potential of LLMs for analyzing web content, it does not assess their ability to act as interactive assistants guiding users through website tasks.

Another relevant area involves the use of LLMs for automated remediation of accessibility barriers. Othman et al. [32] present a case study in which ChatGPT is used to correct WCAG 2.1 accessibility violations detected via WAVE in two non-compliant websites. Their results suggest that LLMs can provide meaningful improvements compared to manual testing alone. This work, however, focuses on code-level repairs, rather than on supporting users in accomplishing tasks on public websites through conversational interaction.

More recently, multimodal LLMs have been studied as copilots for large-scale auditing workflows. Gu et al. [14] propose an AI-assisted framework that operationalizes WCAG-EM by combining a multimodal LLM with graph-based sampling to support human auditors during complex accessibility evaluations. While this demonstrates the value of LLMs in interpreting structural and visual elements of web pages, the focus is on professional auditing rather than end-user support.

Moreover, accessibility evaluation tools that analyze the DOM without relying on LLMs provide context for understanding the challenges associated with automated assessment. Ara and Sik-Lanyi [2] propose a model leveraging structural and visual webpage features to predict accessibility issues across university websites, validated against user testing. These approaches highlight ongoing limitations of rule-based or machine learning evaluators but do not address the interactive and assistive capabilities of conversational agents.

Across these branches of research, no prior work evaluates LLMCAs in terms of their ability to: (1) interpret the structure and interactive elements of real public-sector websites, (2) identify UI components relevant to users' goals and (3) guide users through task completion. In other words, to the best of our knowledge, no existing study examine LLMCAs as real-time accessibility facilitators for users with limited digital skills, cognitive impairments, or other special access needs. The present study fills this gap by conducting a comparative empirical analysis of four popular LLMCAs in real-world navigation scenarios, extending the scope of LLM accessibility research toward user-centred, interaction-oriented assistance.

3 First empirical study: LLMCA capabilities as assistive tool

The first empirical study was conducted by the authors, experts in accessibility and web technologies. The study was performed at the end of 2024 on the homepages of a significant set of popular, widely used public websites, reported in Table 1. As already discussed, at that time the only LLMCA that supported using the web page source code as prompt context without explicitly copying it in the prompt itself was Copilot. Anyway, our aim here was to understand if a LLMCA can perform some specific accessibility tasks, not to compare the LLMCAs performances and potentials (as in the second study reported in Section 4).

3.1 Methodology

We asked Copilot several different questions on each of the pages listed in Table 1. Questions were asked through the Copilot extension in the Edge browser and explicitly instructed the LLMCA to answer using the currently opened web page as the context. Questions belong to two categories:

- Page **structure**: to verify whether the LLMCA can understand the site structure, for example, the location and structure of menus and other parts (or blocks) that make up the web page. Leporini and Paternò [22], in fact, emphasized the importance of splitting the page into logical sections to allow screen reader users to easily overview and navigate the content.
- Page interaction and site **navigation**: to verify whether the LLMCA can understand and explain how to use the site, for example, which button to press to log in or which menu item to select to perform a certain task.

The generated answers were evaluated by the skilled experts to assess their correctness with respect to the page they were applied to. The assessment is based on a 3-level scale:

Table 1 Websites queried in the first study

Website	URL
Amazon	https://www.amazon.it/
Cambridge Dictionary	https://dictionary.cambridge.org/dictionary/english/
Google	https://www.google.com/
Italian Ministry of University and Research (MUR)	https://www.mur.gov.it/
Italian Social Security Service (INPS)	https://www.inps.it/
Poste Italiane (Poste)	https://www.poste.it/
Reddit	https://www.reddit.com/
Repubblica (Italian newspaper)	https://www.repubblica.it/
Tripadvisor	https://www.tripadvisor.it/
University of L'Aquila (Univaq)	https://www.univaq.it/
Weather.com	https://www.weather.com/
Wikipedia	https://www.wikipedia.it/
YouTube	https://www.youtube.com/

- *Correct* answers are completely satisfying, i.e., they give the user all the correct information and directives required by the prompt.
- *Partially* correct answers are considered correct but not exhaustive, i.e., they lack details that would make the interaction more efficient or clearer, or contain additional information not related or not essential with respect to the user question. In other words, this kind of answer is not erroneous but neither is it completely satisfactory.
- *Incorrect* answers are considered erroneous. i.e., the generated information or directives do not accomplish the required task.

However, it is worth noting that, to generate aggregate statistics, we consider both the correct and partially correct answers as "good" ones, since they give the user valuable information related to the question.

Finally, when answers were evaluated as partially correct or even incorrect, we also performed a manual *analysis of the page source* (HTML, embedded style sheets, and scripts) to find a reasonable explanation of the LLMCA misunderstanding or confusion when applied to that page. This has allowed us to make some preliminary hypotheses about why the proposed technique is not applicable in some cases, while results are extremely effective in others.

3.2 Results

The full set of questions and answers, together with further information on the queried sites, is available on GitHub [7]. In the following, we report and discuss some of the most interesting answers, using the notation **[IDn]** to reference the complete data available in the repository above.

Describing the page (structure). In general, we found that the LLMCA can describe the web page contents even in the presence of a complex visual layout if HTML is used in a correct and accessible way. For example, on the *Amazon* website, **[QA1]** *"Can you summarize the product details shown on this page?"* returns a well-written merge between the information given by the "Product Details" page section and an ad-hoc generated summary of the book contents.

Sometimes such descriptions even are smarter than expected, e.g., the question **[QY2]** *"What is the* first *result shown at the top of the page?"*, submitted while browsing a *YouTube* search result page, correctly returns a description of the second result video, (correctly) skipping the first result, which is indeed a sponsored ad. In general, YouTube pages are extremely easy to understand for the LLMCA, despite their apparently complex HTML structure. For example, also questions like **[QY4]** *"What is the rating of this video?"* or **[QY5]** *"Are there any comments on the video? What do they say in general?"* return complete and correct answers.

On pages containing many elements such as boxes, articles, navigation panes, images, etc. arranged in a complex visual layout, the LLMCA, thanks to its knowledge of the page source, can be of help by finding hidden or hardly-visible

page elements. As an example, on the *Repubblica* newspaper homepage, the LLMCA can find and describe, e.g., the main newspaper sections even if they are not visible to the user unless revealed by clicking on some menu or expandable panel. Indeed, [**QR1**] *"What are the main sections of this site?"* correctly lists the items in the "*Sections*" and "*Local editions*" dropdown menus, skipping the "*Services*" one, that actually cannot be considered a newspaper section. On the same site, [**QR3**] *"Can I search for a news on this page?"* describes the position of the search bar, which is actually very small and difficult to locate in such an overcrowded page.

However, the spatial organization of some web pages can sometimes be more *visual* than *structural*, when styles and scripts are used to reposition elements on the page, and this may confuse the LLMCA when "positional" questions are asked. As an example, in a multi-column layout like the *Repubblica* newspaper, [**QR2**] *"What seems to be the most important news published on this page?"* returns a list of news where the "importance" assigned by the LLMCA does not correspond to the one visually deducible from the newspaper interface.

Moreover, the user must be aware that the LLMCA often tries to answer questions even if the requested information is not available on the current page. This can be useful but also misleading in some cases. As an example, on the *Univaq* website, [**QU1**] *"Looking at this page, can you tell me if this university offers a computer science degree?"* gives a correct answer, but the LLMCA here is likely connecting the current page contents (from where it derives the university name) with some external search, since the current page does not contain any information about the requested degree.

Finally, for sake of conciseness (and to save bandwidth), sometimes the LLMCAs tend to "summarize" long answers and therefore must be asked several times to have a complete description of a complex item. As an example, on the *Google* home page, question [**QG1**] *"What kind of task can be accomplished on the currently opened page?"* returns a list of tasks that is clearly incomplete.

Finding and describing navigational structures (structure/navigation/ interaction). Finding and describing menus and other navigational structures accessible from the page is one of the most important tasks when navigational assistance is concerned.

On most modern websites designed for wide accessibility, questions such as *"Can you tell me if there is a menu on the currently open page?"* are correctly handled by the LLMCA. As an example, on the *INPS* site, the answer to question [**QI1**] describes, besides the menu itself, even the *"Go to main menu"* link visible only to screen readers, and the question [**QI2**] *"Can you describe the main menu items to me?"* shows that the LLMCA correctly identifies the main (header menu) items and merged them with the site map (footer menu) items to obtain a complete navigational description of the page.

Sometimes the menus are correctly identified, but the LLMCA does not list all of their items since, as already discussed, it tends to "summarize" long answers. As an example, on the *Wikipedia* homepage, question [**QW3**] *"Are there menus on the page?"* does not return a full description of the menu items, but only the most important ones. The same happens of the *Univaq* website, where [**QU3**] *"Can you tell me if the didattica item in the site top menu has a submenu, and in this case can you list me its sub items?"* returns a not exhaustive list (some subitems are missing), where some item names are even abbreviated.

Since HTML5, there are specific markup elements to mark navigational structures, and the LLMCA analysis clearly understands them. However, in general the LLMCA seems to interpret as menus all the menu-like structures, i.e., nested lists of links with headings. This is an advantage in many cases but can lead to errors in particular web pages. As an example, on the *Tripadvisor* site, the list of items at the top of the page may appear as a menu also to the human reader, but it is actually the category selector for the search box, so it is not surprising that question [**QT2**] *"Are there menus on this page?"* on such a site returns a description of this list as it was a menu, even if it is declared with the "tablist" role in the page source.

However, script-based, dynamically generated menus are hardly detected, confirming that script-based navigational structures are always an obstacle to accessibility. As an example, asking [**QA3**] *"Is there a menu on this page?"* on the *Amazon* home page returns a description of three the menu-like structures present on the page, but misses the overflow left menu, whose structure is not "menu-like" and completely script-driven. The same happens, e.g., on the *Weather.com* web page, where the question [**QW1**] *"Is there any menu on this page?"* returns a description of the top navigation bar only, while the overflow menu is not detected. As an extreme case, on sites like *Reddit*, which uses completely custom tags interpreted through scripts, [**QR1**] *"Is there a menu on this page? Can you list its items?"* shows the confusion of the LLMCA when facing extremely customized structures, returning a mix of internal site links (such as "skip to main content"), search box options, top bar buttons and sidebar items.

Finally, poorly written web pages are also an obstacle for the LLMCA interpretation of the navigational structures. Indeed, on the *Univaq* website, question [**QU10**] *"How can I proceed on this page to enroll to this university?"* returns a strange answer, addressing menus that are not on the page. However, looking at the page source, it is possible to see that such menus are coded, but permanently hidden, probably as a part of a "dead part" of the site structure.

Searching (navigation/interaction). Searching is a very common task on web pages, but finding the search bar and configuring the search may require indications.

On pages with a rich content and a complex visual layout, such as *Tripadvisor*, question [**QT1**] *"What can I search for on this page?"* returns a complete and well-written description of the available search categories and [**QT5**] *"How do I search for a specific restaurant using this page?"* returns complete step-by-step indications on the interface elements to use and how to locate them on the webpage.

The same happens on sites like, e.g., the *Cambridge Dictionary*. Here, question [**QC3**] *"How can I search for a word?"* returns a good description of the search procedure, but when talking about the search bar it only says that is "at the top of the page". Here we can exploit the discursive skills of the LLMCA, asking for further details, such as [**QC4**] *"How can I locate the search bar?"*, which returns a more fine-grained description of the search bar location on the page.

On well-written websites, search directions can be even more detailed. As an example, on the *Poste* website, question [**QP4**] *"Does this page contain a search engine? What can I search for?"* returns the complete list of searchable items, whereas the general (not explicitly search-related) question [**QP7**] *"Is it possible to track the status of a shipment from this page? How?"* directs the user to the correct interaction with the search engine to search for a shipment.

Surprisingly, well-known and apparently simple-looking pages such as *Google* generate only partially correct indications from the LLMCA: indeed, if asked [**QG5**] *"How can I submit a search through this page?"*, the LLMCA tells the user that clicking on the magnifying glass icon submits the search, but this is not true, at least on the search engine homepage. Interestingly, this action is correct on the interface of the search results pages, suggesting that the LLMCA is approximating its answer using some "well known design standard". Moreover, in the LLMCA answer the "Search with Google" button is not cited as an alternative way to submit the search. Then, if asked [**QG8**] *"What are the buttons on this page used for?"* the LLMCA lists many interactive interface items that are not buttons but simply links, and continues to ignore the only elements with a clear button appearance, i.e., those below to the search bar, even if they are clearly marked using the HTML role attribute. Similarly, on the *Weather.com* home page, question [**QW4**] *"How can I search for the forecasts of a specific location using this page?"* gives general and not page-specific indications, which match most of the required interaction, but again tell the user to "click the search icon", which is not clickable.

Searching can be accomplished also through successive refinements of an initial query, and in this case, much more user interaction is needed. Here the LLMCA can be successfully exploited asking for indications on how to apply specific search filters. As an example, on a *Google* results page, the answer to question [**QG10**] *"Can I restrict the search to books only?"* correctly indicates how to locate and use the search category filter, and the same happens with [**QA7**] *"How can I restrict my search to books only?"* asked on

an *Amazon* results page. On the latter site, filtering the results is even more important, so we also checked questions like [**QA8**] *"How can I restrict the search results shown in the current page to English books only?"*, [**QA9**] *"How can I sort the result by price?"* and [**QA10**] *"Is there any way to filter out results with low customer reviews?"*, always obtaining correct indications.

Finally, after searching on *Amazon* the user may try to buy a product. Also in this case, questions like [**QA14**] *"How can I buy this product?"* or [**QA15**] *"Can I purchase more than one copy of the same book?"* always return satisfying indications, covering the interaction with the current webpage as well as the steps to accomplish in the next ones, merging the knowledge deriving from the current page with the directions available in other knowledge bases.

Changing page settings (interaction/accessibility). Other common tasks on web pages are, e.g., switching between available languages or finding a more accessible version of a page. As an example, on *Wikipedia*, question [**QW4**] *"How can I view this page in another language?"* returns a very user-friendly step-by-step explanation of how to accomplish this task by interacting with the interface, and question [**QU6**] *"Has this page accessibility features such as high contrast version or a font size selector?"* on the *Univaq* website returns a correct description of the font size selector available on the page.

However, hardly-to-understand web pages such as *Weather.com* may confuse the LLMCA also in this case: as an example, question [**QW5**] *"Can I switch to another language? How?"*, in this case, returns general directions like "look for the language selection option, usually found at the top or bottom of the page" clearly indicating that the LLMCA has not detected the language selector at the top of the page.

Logging in (interaction). Another common task on web pages is signing in. The LLMCA, if asked about indications, usually gives the correct and informative answers in this case, for example on the *INPS* website ([**QI7**]), on the *Tripadvisor* site ([**QT6**]), on the *YouTube* site ([**QY8**]) or on the *Poste* site ([**QP6**]). Only in some exceptional cases, the LLMCA seems to mix some "design standard" with the real page source, giving inaccurate indications. As an example, [**QA4**] "How can I sign in?" on the *Amazon* home page correctly indicates the *"Hello, sign in."* label to click at the page top, but afterwards it describes a login drop-down that is not actually available.

Finding contact or help with a page (structure). Despite the help of the LLMCA, the user may sometimes need further assistance to interact with a website and its services, so quickly finding contacts on the page is another important task. Again, on highly accessible websites, the LLMCA gives the right answer even if the page appears complex both visually and at HTML level. As an example, [**QP5**] *"Does the page indicate how can I contact Poste Italiane?"* (on the *Poste* website) or [**QI3**] *"Can you tell me if there are any contacts on the open page?"* (on the *INPS* website) return a satisfying answer, which can be however completed by asking more detailed questions like [**QI4**] *"Can you tell me which numbers to call?"*.

On the other hand, in some cases the answers indicate that the LLMCA was unable to clearly identify the contacts, and therefore it tries to answer with useful general indications and links deriving from web searches. As an example, on the *Amazon* website, question [**QA2**] *"How can I proceed to contact the customer service from this page?"* returns generic indications that do not point the user to the real customer service link. However, the answer includes the correct address for the Amazon customer service, so the user may browse it instead of clicking on the current page to reach the customer service.

As a note to be further investigated, when asking on the *Univaq* website [**QU4**] *"Can you tell me if you find any contacts at the bottom of the page?"* we received an answer that includes the correct postal address shown at the bottom of the page, a valid (although not present in the page) email address, but a phone number that does not even correspond to the city of L'Aquila. This may be considered a LLMCA "hallucination", but also the result of the LLMCA being previously trained on a different version of the same web page which conflicted with the provided current source.

Overall, questions about structure were 40, and the correct (completely satisfying) answers were 27 (68%) of them, whereas the other 7 were classified as partially correct, as reported in Table 2 and graphically depicted in Figure 1. Only

Table 2 First study: questions about web page structure

Website	Questions	*Correct*		*Partially*		*Incorrect*	
Amazon	1	1	(100%)	0	(0%)	0	(0%)
Cambridge Dictionary	1	1	(100%)	0	(0%)	0	(0%)
Google	2	1	(50%)	1	(50%)	0	(0%)
INPS	6	4	(67%)	2	(33%)	0	(0%)
MUR	4	3	(75%)	1	(25%)	0	(0%)
Poste	5	4	(80%)	1	(20%)	0	(0%)
Reddit	1	0	(0%)	0	(0%)	1	(100%)
Repubblica	2	1	(50%)	0	(0%)	1	(50%)
Tripadvisor	3	2	(67%)	0	(0%)	1	(33%)
Univaq	6	4	(67%)	1	(17%)	1	(17%)
Weather.com	1	0	(0%)	0	(0%)	1	(100%)
Wikipedia	3	2	(67%)	0	(0%)	1	(33%)
Youtube	5	4	(80%)	1	(20%)	0	(0%)
TOTAL	**40**	**27**	**(68%)**	**7**	**(18%)**	**6**	**(15%)**

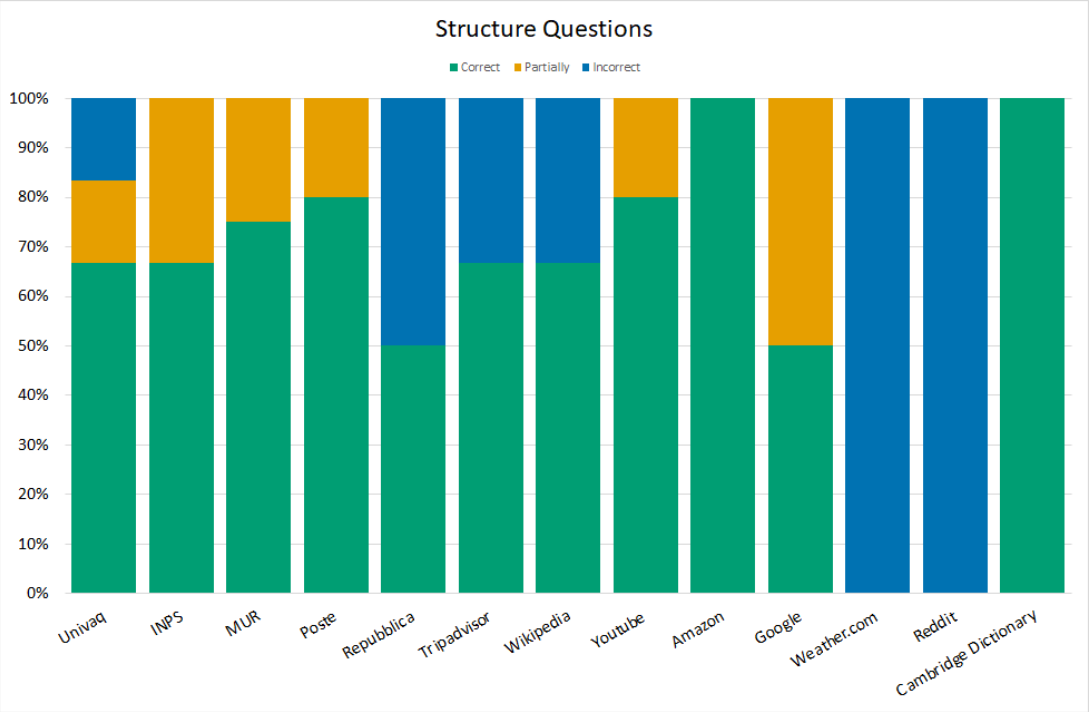


Fig. 1 First study: questions about web page structure. Green bars indicate correct answers, yellow bars indicate partially correct answers, blue bars indicate incorrect answers

in 6 cases (15%) was the returned answer judged completely erroneous (incorrect).

Errors are mainly from a small subset of the queried pages, such as *Reddit*, *Repubblica* or *Univaq*. An analysis of the source code of such pages revealed that they all have a complex HTML structure, i.e., their content is defined using many nested HTML elements and containers, often redundant (mainly to better apply CSS styles) or dynamically manipulated by scripts (to defer loading parts of the content or updated it). Such source complexity is in some cases evident from the webpage appearance (e.g., in sites like *Repubblica* or *Reddit*, which also have a complex visual layout),

but may also remain "hidden" to the website user (for example on the *Univaq* website, which has a rather simple webpage layout). In any case, it represents an obvious obstacle for any algorithm that must interpret the webpage from its source and not from its actual rendering.

On the other hand, the questions related to navigation and interaction were 60, with 38 (63%) completely satisfying and 12 (20%) completely erroneous, as reported in Table 3 and graphically depicted in Figure 2. This time, the manual page source analysis revealed that erroneous answers were mainly from pages that have scarce accessibility, e.g., with heavily script-based navigational structures, sometimes not marked using ARIA accessibility attributes or defined through specific HTML5 markup.

Thus, in both question categories, the (partially or completely) correct answers were above 80%, as shown in Figure 3.

4 Second Empirical Study: assistive LLMCA comparison

Having established that LLMCAs can effectively support users on web page navigation and given the wide availability of RAG and the general evolution of the LLMCA models, at the end of 2025 we performed a further study to compare Copilot with other LLMCAs in our specific accessibility context. In other words, we want to compare different LLMCAs and verify if their RAG capabilities enable them to access the complete source code of a web page, use it to augment the user prompt, and give the user updated and effective answers as Copilot in the first set of experiments.

In particular, the LLMCAs compared were Microsoft *Copilot*, as in the previous study, OpenAI *ChatGPT*, the first and most widely-known LLMCA, *Gemini*, the Google LLMCA also used as predefined assistant on many Android-based smartphones, and *Perplexity*, one of the most challenging competitors of the three "mainstream" LLMCAs above at the time of the experiments.

For this second set of experiments, we used the same list of websites, except for the newspaper website, where we opted for the Wall Street Journal, which is more internationally well-known. The updated list is shown in Table 4.

4.1 Methodology

First, we performed some *prompt engineering* and developed a "setup prompt" to clearly instruct the LLMCA on the kind of answer we expect it to generate and the rules to follow. The prompt requires to talk in English and describes the kind of person the answer must be targeted to. Then it gives some hints on the structure of the expected answers and, most importantly, instructs the chatbot to download the current online version of the specified page and only reason on it while answering the questions.

*Exclusively in this chat session, please consider me *an English-speaking blind person who is not very familiar with using the web*. I need to be guided step by step in interacting with web pages. For each request, you must first provide clear and concise directions based on the visible content of the web page and then provide the same directions in a way that can be accomplished using a screen reader. I will start by giving you the URL of a web page. Please *always download and analyze the current online version of the page*, or alert me if you are using an old or cached version. Your answers must refer *only to the specified web page*: *do not report any information that is not deducible from the page contents and HTML structure*, and *do not refer to any other page in your answers*. Then I will present you with a series of requests, and you will explain to me how to fulfill them on that page - by analyzing its content, interactive elements, and navigation structures - detailing which elements to interact with, where they are located, and how to use them. If the requested information is already available on the page without user interaction, please provide it directly.*

Then we developed a list of questions to ask on each site, trying to maintain a simple, brief and uniform linguistic style. We reused the most interesting queries from the previous study but also added new requests to explore the extended capabilities of current LLMCAs.

Finally, we fed each chatbot with the setup prompt described above, the actual website address to download (as instructed by the setup prompt) and use for the answers, and the list prepared questions. Each question sequence was given in a separate chat, and all the chats were

Table 3 First study: questions about web page navigation and interaction

Website	Questions	*Correct*		*Partially*		*Incorrect*	
Amazon	14	10	(71%)	3	(21%)	1	(7%)
Cambridge Dictionary	6	4	(67%)	2	(33%)	0	(0%)
Google	9	4	(44%)	1	(11%)	4	(44%)
INPS	1	1	(100%)	0	(0%)	0	(0%)
MUR	1	1	(100%)	0	(0%)	0	(0%)
Poste	3	2	(67%)	1	(33%)	0	(0%)
Reddit	4	1	(25%)	0	(0%)	3	(75%)
Repubblica	2	2	(100%)	0	(0%)	0	(0%)
Tripadvisor	3	3	(100%)	0	(0%)	0	(0%)
Univaq	6	4	(67%)	0	(0%)	2	(33%)
Weather.com	4	0	(0%)	2	(50%)	2	(50%)
Wikipedia	3	3	(100%)	0	(0%)	0	(0%)
Youtube	4	3	(75%)	1	(25%)	0	(0%)
TOTAL	**60**	**38**	**(63%)**	**10**	**(17%)**	**12**	**(20%)**

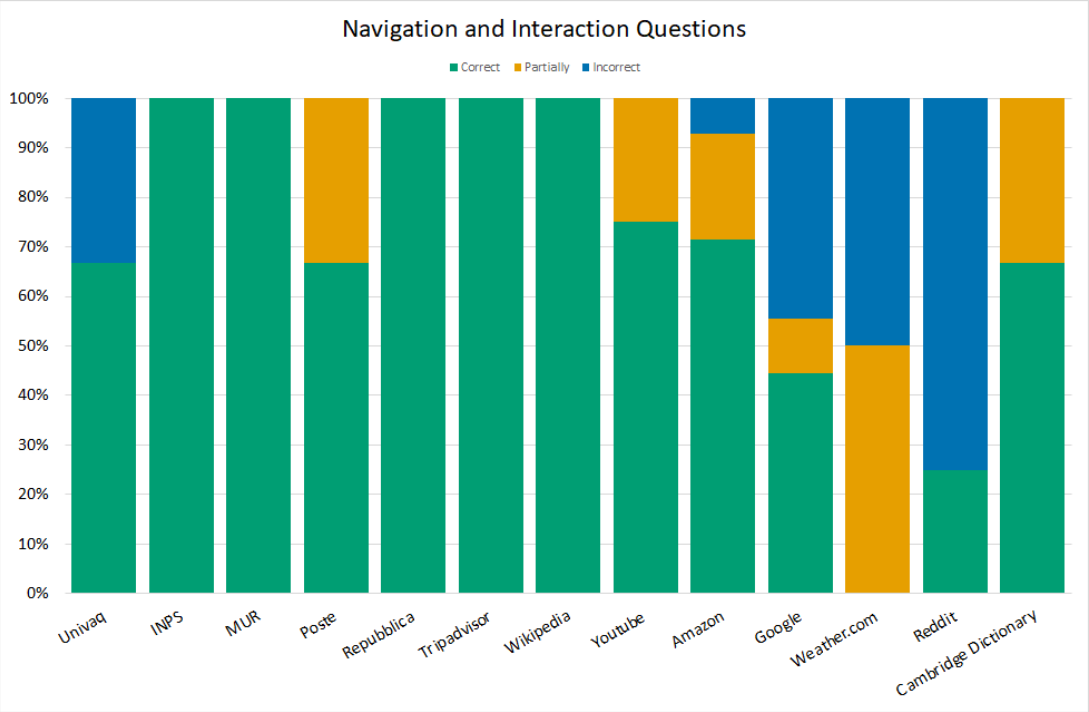


Fig. 2 First study: questions about web page navigation and interaction. Green bars indicate correct answers, yellow bars indicate partially correct answers, blue bars indicate incorrect answers

performed on the free plan of the corresponding chatbot. Note that, when the chatbot was unable to answer due to technical difficulties (fetch blocked, throttled, etc.) we first asked it to describe the problem, and then repeated the experiment three times, on different days. If all the attempts failed, we marked the question as unanswered. Indeed, some sites seem to be blocking most of the requests from certain chatbots, possibly due to data policies.

The complete list of answers was saved, in Markdown format, to separate files named with the question code and the chatbot name (e.g., *QA1_ChatGPT* contains the answer given by



Fig. 3 First study: overall answer quality for questions. Green areas indicate correct answers, yellow areas indicate partially correct answers, blue areas indicate incorrect answers

Table 4 Websites queried in the second study

Website	**URL**
Amazon	https://www.amazon.it/
Cambridge Dictionary	https://dictionary.cambridge.org/dictionary/english/
Google	https://www.google.com/
Italian Ministry of University and Research (MUR)	https://www.mur.gov.it/
Italian Social Security Service (INPS)	https://www.inps.it/
Poste Italiane (Poste)	https://www.poste.it/
Reddit	https://www.reddit.com/
Tripadvisor	https://www.tripadvisor.it/
University of L'Aquila (Univaq)	https://www.univaq.it/
Weather.com	https://www.weather.com/
Wikipedia	https://www.wikipedia.it/
The Wall Street Journal	https://www.wsj.com/
YouTube	https://www.youtube.com/

ChatGPT to the first question on the Amazon website). Both the questions and the answers are available on GitHub [7] where, for each queried website, we also report a saved snapshot taken just before starting the experiments. Note that not everything could be saved in the snapshot for sites with very dynamic structures using JavaScript to load portions of contents asynchronously.

As in the previous experiments, the generated answers were evaluated by skilled experts to assess their correctness with respect to the page they were applied to. This time, we adopted a finer assessment classification based on a 4-level scale:

- 3 = (almost) perfect and satisfying answer: the chatbot answer corresponds to an expert human answer.
- 2 = good, with minor problems: the chatbot answer is acceptable, but less detailed or precise than the expert human one.
- 1 = partially correct: the chatbot answer only partially matches the expert human one, something important is missing or something is clearly wrong.
- 0 = the chatbot was unable to answer (due to repeated technical problems) or the answer was completely wrong.

Moreover, for each answer, we also collected a brief textual evaluation justifying the points assigned to it.

Finally, to further analyze the differences between the four LLMCA answers to the same questions, we also derived a set of objective measures from each of them. In particular, we measured:

- **The BERTscore:** leverages on BERT (Bidirectional Encoder Representations from Transformers) to measure the similarity between two

texts. Unlike other metrics, BERTScore evaluates text based on contextual embeddings and can assess how well a candidate text captures the meaning of the reference text (baseline), considering each word's context. In other words, it allows us to measure *semantic* similarity more effectively.

- **The TF-IDF (Term Frequency-Inverse Document Frequency) Cosine Similarity:** TF-IDF is used in information retrieval to assess the importance of a term with respect to a corpus. The word weight is obtained by multiplying term frequency (TF) in the given document and Inverse Document Frequency (IDF) in the entire corpus. In our case, the TF-IDF vectors, containing the word weights calculated on a corpus of two answers, the candidate and the baseline ones, are then compared using cosine similarity, to assess the similarity of the answers from the most important words they contain.

While such measures can give us some kind of objective data to reason on, they can only assess the lexical and semantic similarities between the candidate answers and the baseline ones, considered as correct or mostly correct. To this aim, our comparison will use as baseline the answers of the overall best LLMCA, as devised by the human-assigned scores above. Indeed, human evaluation, although inherently subjective to some extent, is the main measure to consider, especially considering the specific context we are working on.

4.2 Results

The following Table 5 shows the number of questions asked for each website (86 in total) and the number of answers collected from each LLMCA. When the number of answers is less than the number of questions, the LLMCA explicitly refused to answer or could not answer due to repeated technical difficulties, as explained in the previous sections.

Gemini was the LLMCA with most technical difficulties: it was blocked by three websites, followed by Copilot (blocked by two), and ChatGPT and Perplexity, both unable to query a single (different) website. The most problematic websites were *Amazon* and *WSJ*, possibly due to content access restrictions related to their contents.

Table 6 shows the distribution of points assigned to LLMCA answers, as well as the overall average score of each LLMCA. Gemini shows both the highest percentage of zeroes (26%) and threes (40%). As discussed above, such zero-point answers mainly depend on the sites that blocked the LLMCA, so no answer could be given. When it was able to answer, however, Gemini often returned very detailed and well-structured answers.

Copilot is the second in the three-points ranking, with 35% of "perfect" answers, and has less zero-point answers. Perplexity wins on the 1-points answers, since it often does not respect the "answer from this page only" rule and, in general, gives very generic answers, proving that it is unable to work on the page structure and/or prefers looking for "help" through web searches.

To better compare such values, Figure 4 graphically shows the point distribution.

From Figure 4 Copilot and Gemini are head-to-head as the best LLMCAs in this experiment, whereas Perplexity returned the less useful answers. Indeed, Gemini appears to be, on average (last column of Table 6), the best assistant, with a (narrow) advantage of 0,02 point on Copilot. ChatGPT is third, with 0,07 less than Copilot and the last one is Perplexity, with a big gap of 0,5 with respect to ChatGPT.

As we know from the analysis performed in the first study that some websites have a hard-to-understand dynamic structure, let us drill down to see the average score of the LLMCAs on each website, as shown in Table 7.

Looking at the *Average* columns, *WSJ* was the hardest website, as was *Repubblica* (another newspaper) in the first batch of experiments: a dynamically loaded, crowded interface that is difficult to understand from the source code. *Weather.com* and *INPS* where the easiest to analyze, showing a clear and highly accessibility-enabled (for *INPS*) structure. Anyway, overall average scores are significantly lowered by the presence of the low scores of Perplexity. If we focus only on the best LLMCAs as devised above, e.g., Gemini, we can see that its scores are near or higher than 2 in all cases where the LLMCA was able to answer (value $>$ 0). This indeed shows that, even if not perfect, Gemini answers are on average useful and mostly correct.

Table 5 Second study: overall number of questions and answers

Website	Questions	Answered			
		ChatGPT	Copilot	Perplexity	Gemini
Amazon	6	6	6	0	0
Cambridge	7	7	7	7	7
Google	8	8	8	8	0
INPS	7	7	7	7	7
MUR	6	6	6	6	6
Poste	8	8	8	8	8
Reddit	5	5	0	5	5
Tripadvisor	6	6	6	6	6
Univaq	11	11	11	11	11
Weather	5	5	5	5	5
Wikipedia	6	6	6	6	6
WSJ	5	5	0	5	0
YouTube	6	0	6	6	6

Table 6 Second study: distribution of points assigned to LLMCA answers

Website	Questions	0 points		1 point		2 points		3 points	
ChatGPT	12	(14%)	19	(22%)	36	(42%)	19	(22%)	1,72
Copilot	15	(17%)	18	(21%)	23	(27%)	30	(35%)	1,79
Gemini	22	(26%)	6	(7%)	24	(28%)	34	(40%)	1,81
Perplexity	14	(16%)	43	(50%)	24	(28%)	5	(6%)	1,23

Finally, let's focus on some of the most interesting questions that were asked to the LLMCAs. First, *"Are there menus on this page? Please describe all the menu options"*. This question, that was also asked in the first study, aims to assess if the LLMCA can understand the website code and detect menus and other navigation structures, even if not clearly marked using accessibility tags. It was asked on all the websites except the ones that clearly do not contain menus (e.g., the Google homepage), as shown in Table 8. Note that, for sake of simplicity, in the following tables we exclude Perplexity as it is clearly the worst LLMCA in our experiment. Sites like *INPS* or *Poste*, developed with high accessibility standards, determine almost perfect deductions, even if their menus are complex and often visually hidden to the user (*slide-in*, *script-driven*). Again, if we exclude the non-answer slots (zeroes), Gemini is the best for this task, followed by ChatGPT. Gemini generated optimal menu descriptions even for *Univaq*, where the other LLMCAs nearly failed.

Focusing on another common user task, e.g., logging in. Table 9 shows Gemini and Copilot as the best LLMCAs. In general, however, all the LLMCAs were useful when asked "*Is there a way to log in on this page?*".

Finally, Table 10 shows the statistics for search-related questions (that were site-specific). Gemini and Copilot remain the best. Searching hints are generally acceptable (as the table shows, 2 is the most common score), with some surprising results, i.e., ChatGPT that could not detect the search bar on *INPS*.

To conclude this second set of experiments, having devised that Gemini is, on average, the best candidate assistive tool, we used it as the baseline to calculate the objective metrics described in Section 4.1.

In Table 11 we report the average BERTscore measures on all the answers of each website using Gemini as the baseline.

Blank cells indicate that the LLMCA did not answer for a specific website, whereas completely blank lines indicate sites where the baseline (Gemini) was unable to answer, as explained above. In general, results are very high and close in all cases, indicating that the LLMCA answers seem to be very similar to each other. However, the highest similarity between Gemini and Copilot (which

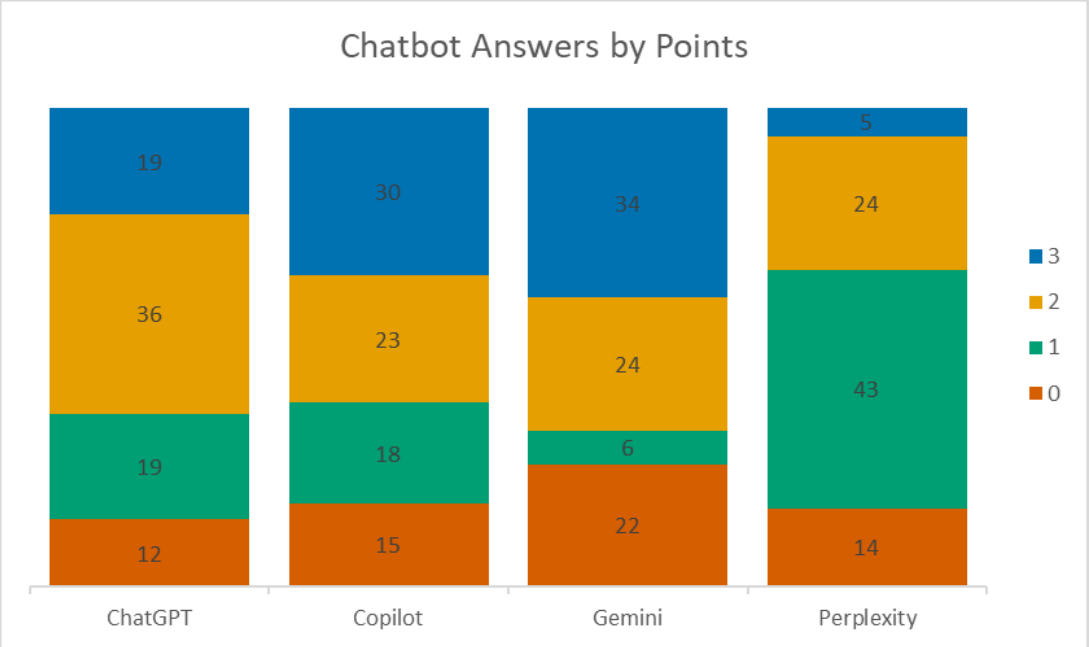


Fig. 4 Second study: overall distribution of points assigned to chatbot answers

Table 7 Second study: average points scored by LLMCAs on each website

Website	**ChatGPT**	**Copilot**	**Perplexity**	**Gemini**	***Overall***
Amazon	1,83	2,83	0,00	0,00	1,17
Cambridge	2,00	2,00	1,43	2,57	2,00
Google	2,00	2,25	1,50	0,00	1,44
INPS	2,00	2,29	1,29	2,43	2,00
MUR	1,67	1,33	1,33	2,33	1,67
Poste	1,75	2,25	1,13	2,63	1,94
Reddit	2,00	0,00	1,00	2,00	1,25
Tripadvisor	2,50	2,00	1,00	1,83	1,83
Univaq	1,18	1,27	1,09	2,64	1,55
Weather	2,60	2,00	1,60	2,20	2,10
Wikipedia	1,83	1,83	1,50	1,83	1,75
WSJ	1,40	0,00	1,40	0,00	0,70
Youtube	0,00	2,67	1,83	2,33	1,71
Overall	1,72	1,79	1,23	1,81	

justifies the overall ranking discussed above) is clear also from these measures, since Copilot (and ChatGPT with a very small gap) has the highest number of top scores (indicated in bold) with respect to Gemini.

On the other hand, while also Perplexity shows high BERTScores, in some cases the highest compared to the baseline, this does not mean it gave responses with comparable quality. Indeed, as already discussed, BERTScore measures how similar the responses are to a reference, but it does not check for accuracy or completeness: responses that are close in meaning to the baseline can still be rated much lower by human reviewers, whose assessments, in our context, are still crucial for detecting issues that automatic metrics overlook.

Table 8 Second study: scores for the question "Are there menus on this page? Please describe all the menu options."

Website	ChatGPT	Copilot	Gemini	*Overall*
Amazon	2	2	0	1,33
Cambridge	1	0	3	1,33
INPS	3	2	2	2,33
MUR	2	0	3	1,67
Poste	3	3	3	3,00
Reddit	2	0	2	1,33
Tripadvisor	3	2	1	2,00
Univaq	1	1	3	1,67
Weather	2	1	1	1,33
Wikipedia	2	2	0	1,33
Average (excl. zeroes)	2,10	1,86	2,25	

Table 9 Second study: scores for the question "Is there a way to log in on this page?"

Website	ChatGPT	Copilot	Gemini	*Overall*
Amazon	2	3	0	1,67
INPS	2	2	3	2,33
MUR	1	3	3	2,33
Poste	2	2	2	2,00
Tripadvisor	2	3	2	2,33
Average (excl. zeroes)	1,8	2,6	2,5	

The calculated TF-IDF values (using Gemini as the baseline), reported in Table 12, substantially confirm the BERTscores discussed above, with some small exceptions.

5 Discussion

The results of the first study, reported in Section 3, suggest that our hypothesis is possible. Answering the first two research questions: AI-powered tools in general recognize the structure of a webpage (RQ1) and understand the interactive elements to support users in performing wished tasks (RQ2). This first study also confirms the value of accessibility for guaranteeing access to all. Indeed, if the HTML is composed in a correct and accessible way, the LLMCA can often describe the web page contents even in the presence of complex visual layouts, and thanks to its ability to interpret the page source, it can detect hidden or hardly visible page elements and deliver smart descriptions that combine the page content with other knowledge sources. On the other hand, script-based navigational structures are an obstacle to accessibility: dynamically generated menus are hardly detected, and poorly written web pages can be an obstacle for the LLMCA interpretation of the navigational structures.

The results of the second study (Section 4) confirm, in general, our deductions about RQ1 and RQ2: the quality of the answers of most of the LLMCAs is high, and well suited to be used as an assistive tool. Moreover, to answer RQ3, we observed that while some LLMCAs report very similar answers in terms of quality and completeness, e.g., Gemini and Copilot, there are cases where some LLMCAs, like Perplexity, have proven less suited to the task. In other cases, LLMCAs were unable to assist the user due to technical restrictions that blocked them from downloading the web page contents, and such restrictions apparently do not apply uniformly to all the LLMCAs.

6 Conclusion

The main objective of our study was to examine the potential of LLMCAs as assistive technologies in the context of web interaction. LLMCAs have been already used to describe (website) interfaces

Table 10 Second study: scores for the search-related question

Website	ChatGPT	Copilot	Gemini	*Overall*
Cambridge	3	3	2	2,67
Univaq	0	0	2	0,67
Weather	2	2	3	2,33
INPS	0	2	2	1,33
Amazon	2	3	0	1,67
Wikipedia	2	2	3	2,33
Tripadvisor	2	2	3	2,33
WSJ	2	0	0	0,67
Google	3	3	0	2,00
Average (excl. zeroes)	2,29	2,43	2,50	

Table 11 Second study: average BERTscore of the LLMCA answers on each website using Gemini as baseline.

Website	ChatGPT	Copilot	Perplexity
Amazon			
Cambridge	**0,858**	0,848	0,849
Google			
INPS	0,821	**0,825**	0,824
MUR	0,822	0,828	**0,832**
Poste	**0,840**	0,838	0,822
Reddit	0,845		0,845
Tripadvisor	**0,855**	0,851	0,846
Univaq	0,833	0,837	**0,839**
Weather	0,853	**0,874**	0,856
Wikipedia	0,847	**0,856**	0,844
WSJ			
Youtube		**0,838**	0,830

Table 12 Second study: average TF-IDF value of the LLMCA answers on each website using Gemini as baseline.

Website	ChatGPT	Copilot	Perplexity
Amazon			
Cambridge	**0,493**	0,490	0,453
Google			
INPS	**0,338**	0,291	0,221
MUR	0,245	**0,377**	0,288
Poste	0,409	**0,434**	0,269
Reddit	0,393		**0,413**
Tripadvisor	**0,482**	0,381	0,447
Univaq	0,327	**0,379**	0,328
Weather	0,467	0,529	**0,551**
Wikipedia	0,461	**0,483**	0,474
WSJ			
Youtube		**0,473**	0,425

as *images*, i.e., working on the visual appearance of the rendered interface. Our idea is to hybridize the current assistive technologies, which mainly work on the web page source (HTML, CSS), with the LLMCA approach by making such sources part of the LLMCA prompt or query context (through RAG). Indeed, LLMCAs are very popular in the development field as code generators and reviewers, so we feel that they should be also able to easily "understand" the meaning and (visual) outcome of extremely complex web pages from their source code and therefore "reason" about their structure and appearance. Clearly, the correct use of well-known web accessibility features, together with a proper use of the web page definition languages, would help the LLMCAs in this task, but their capabilities go beyond such features, making it possible also to apply them on "old-style" websites, which are very hard to approach with other assistive technologies.

In this paper we report two empirical studies that support the feasibility of our approach and explore its possible applications. We asked more than a hundred questions to a selected group of LLMCAs while interacting with a representative sample of widely used websites, such as public administration sites, university sites, and booking services. Questions range from the general description of the webpage contents to the location and description of the navigation elements to the detailed instructions to accomplish a specific goal on a webpage.

Results show a good degree of pertinent and good-quality answers. Moreover, the user-friendly and near-human way to interact with a LLMCA makes it much more comfortable to use for the

average user: questions can be asked in natural language, and answers are usually rich and detailed. This enables this kind of new assistive technology not only to complement existing technologies but also to support or replace the human assistance that is often required for many users to effectively access the web. By eliminating the need to be supported by human help, the user experience of users with disabilities or unskilled web users would be greatly improved, as well as their degree of autonomy, another very important aspect.

The study also highlights that some LLMCAs appear more "skilled" in our accessibility tasks than others. However, such software is currently in continuous and rapid evolution, thus, the results presented in this paper may quickly improve.

Having assessed the feasibility of the proposed methodology in our context, the next natural steps of our research include the involvement of a group of end-users in the study to evaluate non-expert feedback about the LLMCA answers, as well as the extension of our comparative study to new LLMCAs such as Claude by Anthropic, which is quickly becoming the best AI tool for coding, and therefore may be very effective also in our specific accessibility tasks that involve deep code understanding.

Finally, tools of those we investigated, to be effectively exploited as assistive technologies, should be as transparent as possible to the user, and particular attention should also be paid to the user interaction layer, especially regarding the types of requests users are expected to formulate, as well as input modalities such as voice, commands, and gestures. On the other hand, current LLMCA websites often present usability difficulties, particularly in navigation structure, clarity of feedback and interactive elements [23]. Therefore, we intend to design and develop plug-ins or other services that could be embedded within the operating system or browser environment in order to make such specific LLMCAs assistance always readily available to the users.

Declarations

- Funding: No funding was received to assist with the preparation of this manuscript.
- Competing interests: The authors declare no competing interests.
- Data availability: The full set of questions submitted to LLMCAs and the corresponding answers, together with further information on the queried sites, is available on GitHub [7].
- Author contribution: All authors contributed equally to this work.
- Materials availability: Not applicable.
- Code availability: Not applicable.
- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable

References

[1] An J, Yi S, Lyu Y, et al (2025) Conversational agents for older adults' health: A systematic literature review. arXiv preprint arXiv:250323153

[2] Ara J, Sik-Lanyi C (2025) Automated evaluation of accessibility issues of webpage content: tool and evaluation. Scientific Reports 15(1):9516

[3] Bieniek J, Rahouti M, Verma DC (2024) Generative ai in multimodal user interfaces: Trends, challenges, and cross-platform adaptability. arXiv preprint arXiv:241110234

[4] Bonikela HR, Singh N (2025) Predictive ai for web accessibility: Enhancing usability for disabled users. International Journal of Creative Research Thoughts 13(2):410–429

[5] Buzzi MC, Buzzi M, Della Penna G, et al (2023) Accessibility of e-government websites in italy: The user experience of people with disabilities. In: WEBIST, pp 59–70

[6] Chu CH, Nyrup R, Leslie K, et al (2022) Digital ageism: challenges and opportunities in artificial intelligence for older adults. The Gerontologist 62(7):947–955

[7] Della Penna G, Buzzi M, Leporini B (2025) Full Q&A sources for the accessibility studies. GitHub: https://github.com/gdellapenna/AssistiveGenerativeAI

[8] Della Penna G, Buzzi M, Leporini B (2025) Generative AI as a new assistive technology for web interaction. In: IFIP Conference

on Human-Computer Interaction. Springer Nature Switzerland, Cham, Cham, pp 91–107

[9] Deng X, Gu Y, Zheng B, et al (2023) Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems 36:28091–28114

[10] Dombrovskaia L, Vilches A (2021) Guidelines for web pages design inclusive for senior users. In: 2021 40th International Conference of the Chilean Computer Science Society (SCCC), IEEE, pp 1–7

[11] Egard H, Hansson K (2023) The digital society comes sneaking in. an emerging field and its disabling barriers. Disability & Society 38(5):761–775

[12] Ghosh A, Reddy MP, Kodandaram SR, et al (2024) Screen reading enabled by large language models. In: Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility, pp 1–5

[13] González-Sordé M (2025) Exploring the potential of easy language for enhancing website sustainability. Universal Access in the Information Society 24(3):1905–1915

[14] Gu M, Wang Z, Lai S, et al (2026) Towards scalable web accessibility audit with mllms as copilots. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 38515–38523

[15] Hambley A, Yesilada Y, Vigo M, et al (2023) Web structure derived clustering for optimised web accessibility evaluation. In: Proceedings of the ACM Web Conference 2023, pp 1345–1354

[16] Huq SM, Maskeliūnas R, Damaševičius R (2024) Dialogue agents for artificial intelligence-based conversational systems for cognitively disabled: A systematic review. Disability and Rehabilitation: Assistive Technology 19(3):1059–1078

[17] Imtiyaz I, Parvaz M, Mandha S, et al (2025) Assessing RAG: A comprehensive review of evaluation frameworks. In: 2025 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, pp 490–495

[18] Jiang N, Hasanzadeh S, Duffy VG (2025) Domain-tailored generative AI for personalized assistant. In: HCI International 2024 – Late Breaking Papers. Springer, Cham, Cham, pp 233–249

[19] Kamran M, Malik M, Iqbal MW, et al (2022) Web simplification prototype for cognitive disabled users. Human Behavior and Emerging Technologies 2022(1):5817410

[20] Khalil RA, Ahmad K, Ali H (2026) Redefining elderly care with agentic ai: Challenges and opportunities. IEEE Open Journal of the Computer Society 7:326–342. https://doi.org/10.1109/OJCS.2026.3650842

[21] Kim Y, Lee J (2024) Digitally vulnerable populations' use of e-government services: Inclusivity and access. Asia Pacific Journal of Public Administration 46(4):422–446

[22] Leporini B, Paternò F (2008) Applying web usability criteria for vision-impaired users: Does it really improve task performance? International Journal of Human–Computer Interaction 24(1):17–47

[23] Leporini B, Buzzi M, Della Penna G (2025) A preliminary evaluation of generative ai tools for blind users: Usability and screen reader interaction. In: Proceedings of the 18th ACM International Conference on PErvasive Technologies Related to Assistive Environments. Association for Computing Machinery, New York, NY, USA, PETRA '25, p 562–568, https://doi.org/10.1145/3733155.3737910, URL https://doi.org/10.1145/3733155.3737910

[24] Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, pp 9459–9474

[25] Li H, Gai J, Bu L, et al (2025) Enhancing universal access in elderly health information

systems: the triumvirate of trust, adherence, and cognitive load with explainable ai. Universal Access in the Information Society pp 1–15

[26] López-Gil JM, Pereira J (2024) Turning manual web accessibility success criteria into automatic: An LLM-based approach. Universal Access in the Information Society pp 1–16

[27] Lù XH, Kasner Z, Reddy S (2024) WebLINX: Real-world website navigation with multi-turn dialogue. arXiv preprint arXiv:240205930

[28] Martins N, Ralha S, Simoes R (2021) ICT4Silver: Design guidelines for the development of digital interfaces for elderly users. In: International Conference Innovation in Engineering. Springer International Publishing, Cham, Cham, pp 178–188

[29] Mitra DS (2023) Generative AI and metaverse: Companionship and assisted living for elderly people. Available at SSRN 4843464

[30] Moreno L, Petrie H, Martínez P, et al (2024) Designing user interfaces for content simplification aimed at people with cognitive impairments. Universal Access in the Information Society 23(1):99–117

[31] Núñez VJM, Lamas OL, De la Prieta F (2024) Generative AI: Simplifying text for cognitive impairments. In: Methodologies and Intelligent Systems for Technology Enhanced Learning, 14th International Conference. Springer Nature, Cham, p 33

[32] Othman A, Dhouib A, Nasser Al Jabor A (2023) Fostering websites accessibility: A case study on the use of the large language models chatgpt for automatic remediation. In: Proceedings of the 16th international conference on pervasive technologies related to assistive environments, pp 707–713

[33] Patel V, Modi A, Mistry H, et al (2025) From alt-text to real context: Revolutionizing image captioning using the potential of LLM. International Journal of Scientific Research in Computer Science Engineering and Information Technology 11(1):379–387

[34] Pereira L, Brandão D, Martins N (2021) Ageing related human factors to be addressed in the design of visual interfaces of digital applications developed for seniors: A literature review. In: Perspectives on Design and Digital Communication II: Research, Innovations and Best Practices. p 65–80

[35] Pucci E, Possaghi I, Cutrupi CM, et al (2023) Defining patterns for a conversational web. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp 1–17

[36] Raihan MM, Subroto S, Chowdhury N, et al (2024) Dimensions and barriers for digital (in)equity and digital divide: A systematic integrative review. Digital Transformation and Society

[37] Schoenborn NL, Chae K, Massare J, et al (2025) Perspectives on ai and novel technologies among older adults, clinicians, payers, investors, and developers. JAMA network open 8(4):e253316

[38] Shridhar M, Thomason J, Gordon D, et al (2020) ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10740–10749

[39] Silva JSR, Cardoso PCF, De Bettio RW, et al (2024) In-page navigation aids for screenreader users with automatic topicalisation and labelling. ACM Transactions on Accessible Computing 17(2):1–45

[40] The Easy Read Standard (2026) An evidence-based framework for accessible information. https://easyreadstandard.org/

[41] Uricchio T, Ceccacci S, D'Angelo I, et al (2024) Investigating OpenAI's ChatGPT capabilities to improve accessibility of textual information: An explorative study. In: International Conference on Human-Computer Interaction. Springer Nature Switzerland,

Cham, Cham, pp 326–336

[42] Wang C (2024) Artificial intelligence helps the elderly to integrate into the internet age. In: 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024). Atlantis Press, Amsterdam, pp 533–541

[43] Yao S, Chen H, Yang J, et al (2022) WebShop: Towards scalable real-world web interaction with grounded language agents. Advances in Neural Information Processing Systems 35:20744–20757

[44] Yau Y, Shen YC (2025) "though helpful, still hesitant": a tam-based qualitative study on older adults' ambivalent acceptance and model extensions in ai fitness coaches. Frontiers in Psychology 16:1666755

[45] Zhou S, Xu FF, Zhu H, et al (2023) WebArena: A realistic web environment for building autonomous agents. arXiv preprint arXiv:230713854