David Mischak, MA BSc

Context-Based Image Accessibility For Blind People Using Large Language Models

Master's Thesis

to achieve the university degree of

Dipl.-Ing./Master of Science

Master's degree programme: Software Engineering and Management

submitted to

Graz University of Technology

Supervisor

Assoc. Prof. Dr. tech. Dipl.Ing. Johanna Pirker, BSc

Co-Supervisor

Dipl.Ing. Michael Holly, BSc

Institute of Human-Centred Computing
Head: Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Graz, January 2025

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date   Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum   Unterschrift

Abstract

Today, the Internet is part of everyday life, playing a central role e.g. in education, work and entertainment. Blind people who rely on assistive technology to navigate the Internet because of their visual impairment often face barriers. They face particular hurdles when websites contain visual content, such as images, for which there are no or inadequate descriptions, thus they get excluded from these information. Large language models that can analyze images and generate descriptions from them could provide potential solutions to this problem. In addition, prompts can be used to tell the generative AI which image details to highlight.

In the course of the present research project, a literary research was first carried out and then a prototype was developed, that was then evaluated with the aid of expert interviews. In the theoretical research, on the one hand, accessibility measures were discussed that contribute to the inclusion of blind people on the web, and on the other hand, navigation patterns and image perceptions of sighted and blind people were compared in order to identify barriers. Furthermore, the basics of Artificial Intelligence were described. These include text analysis using Natural Language Processing, Generative AI, Large Language Models, optimization options for prompts, and potential risks when using AI. Based on those findings, a prototype was developed that uses Generative AI to describe images on Wikipedia articles for blind people based on the context. This prototype was then evaluated using an empirical qualitative study. For this purpose, interviews with five blind experts were conducted, evaluated based on Mayring and discussed.

Kurzfassung

Heuzutage ist das Internet Teil des täglichen Lebens und spielt under anderem zentrale Rolle in den Bereichen Bildung, Arbeit oder Unterhaltung. Blinde Menschen, die aufgrund ihres fehlenden bzw. zu geringen Sehvermögens beim Navigieren im Internet auf assistive Technologien angewiesen sind, erfahren dabei häufig Barrieren. Besondere Hindernisse entstehen für sie, wenn sich visuelle Inhalte wie Bilder auf Websites befinden, aber Bildbeschreibungen entweder nicht oder nur unzureichend vorhanden sind, sodass sie von diesen Informationen exkludiert werden. Potenzielle Lösungen für diese Problematik könnten Large Language Models bieten, welche in der Lage sind, Bilder zu analysieren und daraus Beschreibungen zu generieren. Zusätzlich kann mittels Prompts der Generative AI mitgeteilt werden, auf welche Bilddetails der Fokus gelegt werden soll.

Im Zuge des vorliegenden Forschungsvorhabens wurde zuerst eine Literaturrecherche durchgeführt und dann ein Prototyp entwickelt, der anschließend mithilfe von Experten:innen-Interviews evaluiert wurde. In der theoretischen Recherche wurden einerseits Accessibility Maßnahmen erörtert, die zur Inklusion blinder Menschen im Web beitragen, andererseits wurden Navigationsmuster und Bildwahrnehmungen von sehenden und blinde Menschen verglichen, um Barrieren zu identifizieren. Des Weiteren wurden Grundlagen zur Künstlichen Intelligenz beschrieben. Diese umfassen die Textanalyse durch Natural Language Processing, Generative AI, Large Language Models, Optimierungsmöglichkeiten von Prompts sowie potenzielle Rikisien bei der Verwendung von AI. Anschließend wurde auf Basis der gewonnen Erkenntnisse ein Prototyp entwickelt, der Bilder auf Wikipedia für blinde Menschen mithilfe von Generative AI kontextbasiert beschreibt, welcher dann mithilfe einer empirischen qualitativen Studie evaluiert wurde. Dafür wurden Interviews mit fünf blinden Expert:innen durchgeführt, nach Mayring ausgewertet und diskutiert.

Acknowledgements

I would like to thank everyone who supported me during my time at the Graz University of Technology and while writing my master's thesis.

First of all, I would like to sincerely thank Assoc. Prof. Dr. tech. Dipl.Ing. Johanna Pirker, BSc for her support and important advice in every phase of the work and for giving me the opportunity to research in the field of accessibility.

I would also like to thank my co-supervisor, Dipl.-Ing. Michael Holly, BSc, who provided me with feedback and advice during the final phase of my thesis.

Thanks also belong to Mag.phil. Barbara Levc, for helping me to find a topic and for conducting a comprehensive interview that allowed me to delve deeper into the research. I would also like to thank David Wohlhart for the insightful discussions and for his support with the research content.

I would also like to express my special thanks to the five interviewees who agreed to take part in the evaluation of the prototype, to be interviewed and were very generous in sharing their experiences, needs, opinions and desires. In this context, I would also like to thank Abigail Faffelberger, BEd for helping to find some of the participants.

Many thanks also to my family, who supported me throughout my studies and especially during the master's thesis period. I would also like to warmly thank my partner Klara Schöffmann, MA MEd for the insightful scientific discussions and advice during the entire period of writing the thesis.

Contents

List of Figures

List of Tables

1. Introduction

We are currently living in the age of digitization with the internet being indispensable across nearly all aspects of daily life, from education and health care to employment and entertainment. The internet is accessed by people with and without disabilities, but people with disabilities often encounter barriers. Blind individuals, on whom this work will focus, are using screen readers to navigate and interpret their content. Web technologies are advancing rapidly, providing developers with countless opportunities to create dynamic and interactive layouts. However, these advancements come with a downside: traditional access methods for blind users often cannot keep pace with these complexities, leading to new barriers (Sulong & Yusof, 2024). As Oyewole (2019) notes, these barriers are frequently unintentional, stemming from a lack of awareness among developers about how to ensure website accessibility. Consequently, blind users frequently face issues on websites such as a incorrect information, disorientation and irrelevant content such as advertisements and unrelated news (Sulong & Yusof, 2024).

A particular challenge for blind individuals arises from the use of images on websites. While sighted users can effortlessly focus on any desired feature of an image using their visual abilities, blind users rely on image descriptions, which are read aloud by screen readers (Firth, 2024). However, these descriptions are not always available or adequately authored. Furthermore, even when such descriptions exist, blind users have no control over the content provided, as the descriptions are written by developers (Stangl et al., 2020).

Challenges like those faced by blind users in perceiving visual content online can be potentially addressed by Generative AI, due to its advanced language and image processing capabilities (Foster, 2023). In the summer of 2018, the company OpenAI introduced GPT-1 laying the foundation for what is now known as ChatGPT. This service was the first version of Generative AI that was accessible and intuitively usable for the global public, so people had a first opportunity to interact with this technology. When people recognized the potential and impact of text-generating AI, their interests in Large Language Models (LLMs) immediately increased (Marr, 2023). Due to the nowadays global availability of these tools, they offer a great opportunity for research in different areas. In the context of this work, this raises the question whether Large Language Models could be applied to generate alternative texts for images. In addition to that, they should also only provide information for blind people, that are actually relevant for them with regard to the context of corresponding websites.

1.1. Goals and Motivation

In 2020, I decided to attend the Master's program in Software Engineering and Management in addition to my first Master's program in Interaction Design, in order to combine the competences of both areas, that are user experience design and software engineering. My personal focus and interest lies mainly on people who use digital products. I always ask myself whether the functionality of the software meets the needs of the user, if the software's user interface is intuitive enough and to which extend the product has a positive impact on them. In this context, a personal central aspect is that solution approaches can be thought out outside the box, in order to solve problems in an unconventional way if necessary.

In the course of my research for the topic of this master's thesis, I came in touch with a blind person who told me about their experiences with screen readers on websites. Blind people face various barriers, such as hurdles while navigating on websites and, in particular, the lack of accessibility of visual content. This problem sparked my interest in learning about the perspective of blind users on websites and using my abilities to help them overcoming these barriers by applying state-of-the-art technologies.

In this sense, the goal of this thesis is to develop a functional prototype that describes images on websites in a context-sensitive way for blind people. On the one hand, an aspect involves determining how Generative AI can be optimally applied to create image descriptions of the highest possible quality. On the other hand, the target is to determine factors such as performance, costs and security of the technologies used in order to implement the prototype.

1.2. Methodology

In order to achieve this research goal, a systematic literature research was first conducted on the topics of blindness, accessibility of websites and images, and artificial intelligence. For this purpose, sources from scientific books, studies, journals, and websites were gathered. In addition, an extensive interview was conducted with the blind expert Mag. phil. Barbara Levc.

Based on the theoretical findings and available technical possibilities, as well as taking into account the private financial resources, a prototype was developed using web technologies, such as a Google Chrome Extension, the Express.js framework and Large Language Models. An empirical and iterative workflow was chosen for this implementation.

Finally, the developed prototype was tested through five expert interviews to consider its usefulness and usability. The results were then presented and reflected upon in a discussion together with the research objective.

1.3. Structure

This thesis is divided into four main parts: the theoretical part (Chapter 2), the practical part (Chapter 3.6), the empirical qualitative evaluation (Chapter 4) and the last part containing lessons learned, future suggestions and a conclusion.

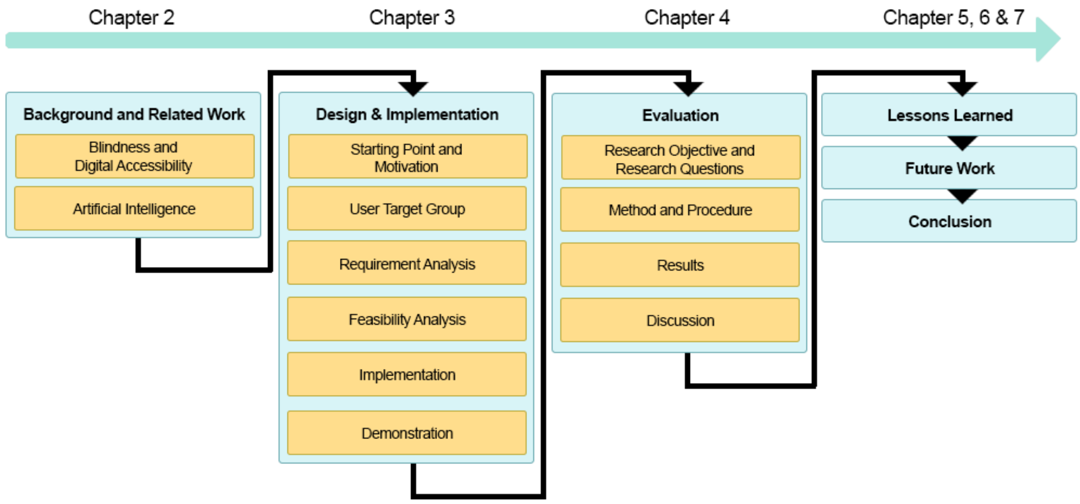

Figure 1.1.: Thesis structure.

In the second chapter, the topics of blindness and web and artificial intelligence are elaborated. In order to get a better insight into blindness and web, the first section deals with the definition of blindness, required measures to include blind people to the web and a comparison between blind and sighted people to identify barriers. Thereon, the second Section covers basics of Artificial Intelligence (AI) with the emphasis on text-generating models. This involves the importance of Natural Language Processing and its recent advancements, special capabilities of Generative AI, the operation principle and versions of

Large Language Models as well as strategies to optimize prompts and general concerns about this technologies.

The third chapter documents the empirical and iterative development process of the prototype. The product idea, the requirements analysis and a series of different experiments are presented. In the course of this, factors of AI such as performance, quality, costs and safety are examined.

The evaluation (fourth chapter) of the developed prototype takes place in the third chapter. For this purpose, interviews with blind experts are conducted using a qualitative research method. The study design is described in terms of methodology, research instrument, participants and transcription method. Subsequently, the results are evaluated on the basis of the empirical qualitative content analysis according to Mayring.

This is followed by the chapters Lessons Learned, Future Work and Conclusion.

In the appendix there is a transcript and translation of the interview with Mag. phil. Barbara Levc, the interview guideline for the qualitative study and the declaration of consent.

2. Background and Related Work

The purpose of this chapter is to provide fundamental knowledge on the fields of blindness, inclusion, accessibility, assistive technologies and Artificial Intelligence. The first subsection (2.1) describes the definition of blindness, factors that cause blindness and existing solutions and concepts, that enable these people to participate in today's society, especially in web. Therefore, universal design, usability, accessibility and assistive technologies will be described. In addition to that, the interaction behavior of sighted and blind users on websites and, in particular, the perception of images are compared from a practical point of view, to understand which procedures work flawlessly and where difficulties arise.

The second subsection (2.2) covers fundamental knowledge about Artificial Intelligence. This involves a definition of AI, giving an introduction to Natural Language Processing (NLP), providing an overview about various Large Language Models, discussing strategies to improve prompts and mentioning concerns from society. This subsection is especially relevant as it forms the basis for the practical implementation of the prototype.

2.1. Blindness and Digital Accessibility

According to estimations of the World Health Organization (2023) (WHO) about two billion people in the world suffer from vision impairments or blindness. Blindness and vision impairments are ranked as the globally third leading disability. Approximately 90% of the affected people live in low to middle income countries where women and ethnic minorities are the most suffering among these (Forrest et al., 2023). However, the exact number of visually impaired people is hard to determine, since data are missing or inaccurate, but most probably the real numbers are greater than numbers provided by statistics (Lang et al., 2020). A lack of data particularly exists in the Caribbean, Central Asia, Latin America and Central Sub-Saharan Africa that makes these regions non prestigious in the global statistics (Forrest et al., 2023). Nevertheless, available data from Austria show that 3.4% of the population suffers from blindness or visual impairments (BSVÖ, n.d.).

Nobody is guaranteed not to develop any eye deceases in lifetime, whereby most severe visual impairments occur over an age of 50 years (World Health Organization, 2023). Data from Germany report that in the year 2017 79% of people with a severe eye condition were older than 60 years (Lang et al., 2020). Thus, the longer people live the more likely they are getting at least one eye condition (World Health Organization, 2023). Due to quickly increasing life expectancy and a decreasing number of child mortality, greater financial investments are urgently required to treat and cure eye conditions. In fact, the large number of visually impaired and blind people causes serious economic and societal challenges (Forrest et al., 2023). Represented in numbers, the annual financial loss due to less productivity caused by barriers, based on visual impairments, encompasses more than $411 billion worldwide (World Health Organization, 2023). As a consequence, accessibility for blind and visually impaired people has to be improved.

In the next section, the terms visual impairment and blindness are explained and the causes and effects are discussed.

2.1.1. Blindness

Blindness an visual impairments are terms most people feel familiar with, however providing a precise definition is relatively difficult to achieve. There are many approaches on defining blindness and listing all of them would exceed the scope of this section, so the most relevant ones will be discussed. The visual system, roughly explained as light that enters the eye and falls onto the retina, going through the optic nerve into the brain, conquering many complex anatomical structures like the visual cortex and resulting in cognitive representations of objects. Due to this complexity, the visual system has a great potential to fail what can lead to vision impairments or blindness (Ray et al., 2016). The five top ranked causes of such lesions are cataract, diabetic retinopathy, glaucoma, age-related macular degeneration and refractive errors (World Health Organization, 2023).

Cataract and glaucoma are deceases that are caused by for example metabolic disorders, genetics, age-related changes and result in problems with accommodation, lack of contrast sensibility, disrupted color vision and further on. Moreover, diabetic retinopathy and macular degeneration are caused by genetics as well as developmental disorders, tumors, toxic lesions, degenerative retinal processes and many more. Consequently, affected individuals can have a bad sight at night, problems with contrasts and vision acuity, defective light-dark adaption and so on (Walthes, 2022). Refractive errors occur when the shape of the eyeball either becomes too long, too short, the cornea distorted or the lens aged, thus it remains fogged. Symptoms of refractive errors are e.g. double vision, headaches and blurry vision when looking at near distance objects (National Eye Institute, 2023).

Generally, different organizations and countries specified distinct methods of measuring the degree of a person's eye condition (Ray et al., 2016). Walthes (2022) describes the term vision impairment with the sub categories blindness, vision disability and severe vision disability. A common way is to describe blindness as a vision impairment that lies under a certain threshold that varies between contexts (Ray et al., 2016). The used measurement units are the visual acuity[1], after a best possible refractive correction (Walthes, 2022), and the size of the visual field[2] (Ray et al., 2016). According to the WHO the visual acuity must be under a threshold of 20/500 and/or the visual field must be less than 10 degrees. The US dictates a visual acuity of under 20/200 and/or a visual field that falls under 20 degrees, whilst the UK only investigates the visual acuity that has to be below 20/400 (Ray et al., 2016). In Germany a person is considered as blind if he or she has a worse visual acuity than 20/1000 on the better eye (Walthes, 2022).

However, it is important to mention, that these classifications do not provide proper insights for finding better solutions, since people are individual and they have different prior experiences that affect their vision impairments diversely (Walthes, 2022). Hyvärinen (2001, cited in Walthes, 2022) suggests, orientated on the International Classification of Functioning, Disability and Health (ICF)[3], that a classification of blindness has to be performed depending on people's activities such as communication and interaction, orientation and movement, daily routine activities and tasks like reading and writing (Walthes, 2022). This definition could be beneficial for a more precise categorization, but the downside is that monitoring of people's lives would be much more complicated (Walthes, 2022).

"The American Foundation of the Blind" made a list of described definitions for various terms of blindness or vision impairment. These are functional limitation, legal blindness, low vision, self-reported vision loss, total blindness, vision difficulty and vision loss. Two presented types of these are useful for this thesis, namely low vision and total blindness. Low vision refers to having some remaining sight, however it is too limited for executing normal tasks, and total

[1]Visual acuity describes the sharpness of the vision. The top number stands for the distance (feet) from which the examined person can read the eye chart correctly, whereas the bottom number describes the distance (feet) of people with normal vision to see the chart accurately (e.g. 20/40) (Porter, 2022).

[2]The visual field describes the amount of vision a person has along the edges of their eyes (Turbert, 2022).

[3]The International Classification of Functioning, Disability and Health (ICF), published by the WHO, is a classification about ability, disability and health. It includes a bio-, psycho- and social perspective (World Health Organisation, 2024).

blindness means to be unable to perceive any visual representation through the eyes (American Foundation for the Blind, 2024). These types of blindness were chosen for this thesis, because the focus is on people who cannot use their sight, so they have to rely on other senses[4].

2.1.2. Web Inclusion for Blind People

As discussed in the beginning, over two billion people in the world suffer from blindness (World Health Organization, 2023). To guarantee the rights of persons with disabilities, including the rights of blind people, the United Nations Convention on the Rights of Persons with Disabilities (Disability Rights Convention, UN CRPD) was adopted in 2006 and opened for signature in 2007 (Austria signed in 2008) (United Nations, 2012).

The Disability Rights Convention is an international agreement in which the signatory states committed themselves to promote, protect and guarantee the human rights of people with disabilities. They must enjoy all human rights and fundamental freedoms. Thus, people with disabilities must be guaranteed the full access to different spheres of life e.g. accessibility (Article 9), living independently and being included in the community (Article 19), education (Article 24), as well as work and employment (Article 27) (United Nations, 2012).

The Austrian report "Report on the implementation of the UN Convention on the Rights of Persons with Disabilities", published in 2018 by the Austrian Disability Council (Österreichischer Behindertenrat), described that the situation for people with disabilities, including blind and visually impaired people, has barely changed since the signing of the UN CRPD in 2008. In some parts of their lives, e.g. education, work and accessibility, conditions have even become worse. This was caused by financial reductions for health care, education, scientific research and social services for people with disabilities (Österreichischer Behindertenrat, 2018).

In order to meet the requirements of the UN CRPD, certain concepts and methods can be applied to remove barriers that blind and visually impaired people face on websites in their everyday lives. There are two types of approaches for achieving accessibility:

1. **Proactive:** Developers already take aspects of inclusion into account when developing their websites (e.g. Universal Design, Usability and Accessibility).

[4] Regarding the other types of vision impairments, which were previously mentioned, their definitions can be found in the literature (see American Foundation for the Blind (2024)).

2. **Reactive:** Blind and visually impaired people use assistive technologies in order to get full access to websites (e.g. screen readers).

Both approaches must work together to ensure full accessibility. A selection of the most important methods/tools are described in the following section.

2.1.2.1. Universal Design

Universal design means that web developers design a product that can be used by the widest possible range of people without requiring an adaption for certain use cases. This design concept involves making a product useful for any kind of abilities, moreover, it must be customizable to individual users requirements. In addition, the design must be intuitive and self-explanatory, so that people with different prior experiences can use and understand provided information efficiently, hence the intuitive design must also ensure users to return to their desired path, even if they accidently clicked on a wrong button or got trapped anywhere. The physical characteristics of a person such as vision and hearing ability must not have any negative impact on the user experience. Consequently, the mentioned aspects should lead to a better social inclusion and equality, however implementing these strategies is often an exhausting challenge for designers, entrepreneurs and more. All in all, there is no single generic approach for making a product usable by everyone and only addressing barriers will not solve the problem (Persson et al., 2015).

The Austrian Working Group for Rehabilitation (Österreichische Arbeitsgemeinschaft für Rehabilitation) claims that products, environments, programs and services must be designed in a way, that all people, including people with disabilities, can use them to a wide extend without any specific adjustments. This does not mean, that universal design makes assistive technologies unnecessary, but it reduces the need of them (Österreichische Arbeitsgemeinschaft für Rehabilitation, 2010).

Usability and Accessibility offer methods to implement universal design, so user-friendly access for websites can be guaranteed for all people. The next sections of the thesis describe the just mentioned approaches.

2.1.2.2. Usability

Usability refers to designing websites, or more specifically user interfaces, in a way to make visitors feel comfortable (Henry et al., 2016). This is essential to ensure that users stay engaged during their visit, so they do not immediately leave the site, that is obviously an important resource for them. A set of usability principles serve as a measure to consider how usable a website actually is. These principles are used during the whole design process and are explained in the following (Nielsen, 2012):

- **Learnability:** How difficult is it to achieve a certain goal without having visited the website before?
- **Efficiency:** After the successful orientation on the new website, how fast can actions be performed?
- **Memorability:** If the website was used in the past before and is revisited again, how fast can users repeat previously learned actions?
- **Errors:** How often do people click on wrong buttons etc. and how fast can they return to their initial path?
- **Satisfaction:** Is the overall layout with all the elements (design) comfortable and would users visit the website again?

In order to ensure a best possible usability for a website, it is important to understand that users show individual behavioral patterns when it comes to interacting with a website. They choose different navigation paths to find the desired information and always try to proceed as quickly as possible. This means, for example, that users do not read all the content of a webpage in detail, but jump from element to element and read in more detail as soon as it seems relevant to them (Hustak & Krejcar, 2016). The next section deals with concepts and methods how efficiency can be applied to users with diverse abilities.

2.1.2.3. Accessibility

While usability deals with creating an effective, efficient and satisfying design for all users, accessibility refers to removing excluding factors that prevent disabled users from using a website (Henry et al., 2016). Therefore, it is essential that the web is equally accessible to all people, regardless of their abilities. In detail this means that these people should be able to perceive, understand and interact with a website. The form of disability can be auditory, cognitive, neurological, physical, linguistic or visual (Henry, 2024). In this regard, the Web Accessibility Initiative (WAI) presents the following statement:

> *"The Web is fundamentally designed to work for all people, whatever their hardware, software, language, location, or ability. When the Web meets this goal, it is accessible to people with a diverse range of hearing, movement, sight, and cognitive ability"* (Henry, 2024)

In the context of blind people, they can access websites via alternative modalities without having to see the visual content. For this purpose screen readers are commonly used, in order to read out loud the content of websites (Firth, 2024).

In order to better understand the technical requirements for websites for screen readers, it is first important to discuss how websites are constructed. A website is built with various HTML tags. Originally, web developers used so called Native HTML tags, that are e.g. button tags, anchor tags, headline tags or image tags. However, over the years, due to evolving technologies and the option to implement more complex designs, developers started to use two categories of elements. There is the category of interactive tags, such as buttons, input fields, check boxes or radio buttons. Another category is layout elements, such as container tags, that are used to put content together in specific layouts. They can be implemented in combination with stylesheets, which gives the layout a certain visual appearance (colors, fonts, text alignment etc.) (Firth, 2024).

All HTML elements are arranged in a hierarchical "tree" structure, the so-called Document Object Model (DOM) tree (see Figure 2.1). The DOM tree is essential for a website to work properly in the web browser, but a sighted user is mostly unaware of the exact layout code, because only seeing the content is relevant for them. Blind people also only care about the content instead of the layout code, however they cannot perceive the visual appearance of the website and screen readers can have problems with handling the layout code properly (Firth, 2024). Mag. phil. Barbara Levc, an expert in the topic of blindness, who was interviewed on the 4th of July[5], confirms the described difficulties in the following quote:

> *"[...] It is different with buttons, so you cannot use the tab key to jump from one button to the next. While a link can always be activated with a screen reader, there may be buttons that do not work. Or which only work in combination with a certain browser, for example, or which do not say that they are not activated, but are activated anyway. So there are a few variations of buttons that do not work 100% barrier-free. There must be something in the background, in the way they are programmed, that is the cause. [...]"* (Levc, 2024)[6]

In order to provide screen reader users with a convenient access to a website's content, there is another form of tree structure, namely the accessibility tree. The accessibility tree (see Figure 2.1) modifies only those parts of the DOM

[5]In order to gain a better insight into the topic of blindness Barbara Levc was interviewed. She works at the University of Graz and is the head of "Zentrum Integriert Studieren", which is responsible for the realization that people with disabilities at the university have equal rights. According to her own words, she suffers from a progressive retinal disease. The medical term for this is tubular vision, after which the cells of the retina stop functioning properly. She described how she could actually see perfectly when she was young, but the progression gradually led to blindness. It started with visual impairment but ended with blindness. Levc can still remember visual perceptions that she could see, but she can only guess the appearance of things she has never seen.

[6]The original interview is attached in the appendix A.

tree, that are relevant for a blind person to be accessed. As a result, by using an accessibility tree a website is understandable and interpretable for screen readers (Firth, 2024).

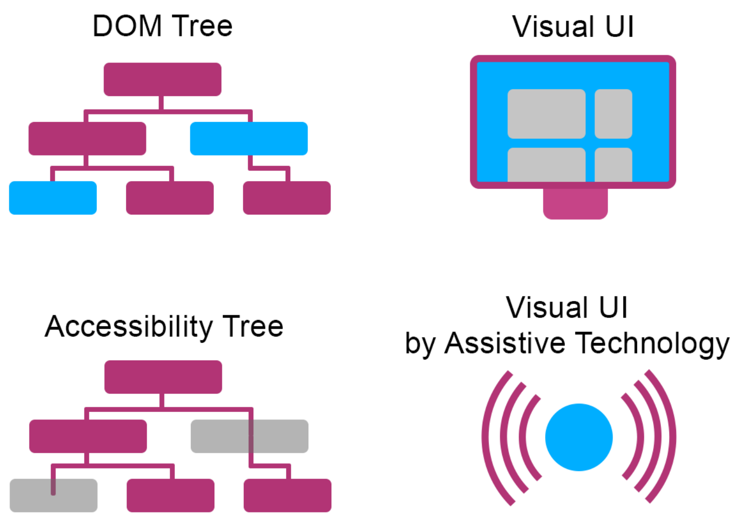

Figure 2.1.: The application and outcome of DOM-Tree and Accessibility-Tree. Based on Firth (2024).

While blind users navigate through websites, by traversing its accessibility trees with their screen readers, the screen readers read out loud information about any focused element (the element that is currently selected). At this point, Native HTML elements are particularly useful, since accessibility features are already included in them. Therefore, these HTML tags provide four attributes: a role, a description, a state and a name (see Figure 2.2) (Firth, 2024).

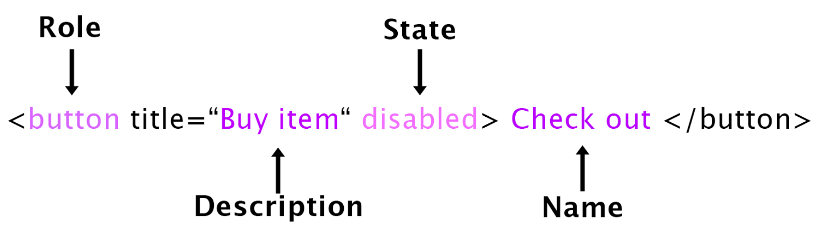

Figure 2.2.: A Native HTML tag with Role, Description, State and Name. Based on Firth (2024).

These properties are listed and briefly described below (Firth, 2024):

- **Role:** This property gives information about the type of an element. This could be a button, an input field, a link, an image and so on.
- **Description:** Its purpose is to provide additional details about what an element is supposed to do, for instance, when referring to Figure 2.2, a button could sign in users into their accounts.
- **State:** The state defines an element's appearance or behavior at a certain point in time. As a result, an input field or a button could be grayed out (disabled), so it is not interactive and normally looks grayish. This look and behavior can change at any time and make an element return to its default color and interactivity again.
- **Name:** This is a text that labels an element. As an example, a button could be labeled "Sign in", as shown in the Figure above.

As already mentioned, due to the changing demands on designs today, more and more developers are for example using container tags, which were considered as layout elements before, to build new and creative interactive elements. The reason is that these HTML elements have a broader range of possibilities to be modified than the native ones. By doing so, they are getting styled and programmed in a way, that they visually appear and behave like a better version of a native interactive HTML tag, and this works for sighted users. However, they are inaccessible for screen readers, because they do not naturally provide any information what the screen reader would require (Firth, 2024). For that reason it is recommended to use Native HTML tags as often as possible.

Another barrier that arises for screen reader users is the accessibility of images, that are shown on websites, because they normally require sight to be perceived. Since blind people cannot capture them by their sight, they are dependent on alternative texts (alt texts), that can be read out loud by screen readers. Alternative texts are supposed to describe content of images. Generally, even though Native HTML works well with screen readers, there is a specific challenge when it comes to image tags. The role of these tags can be easily recognized by a screen reader, thus it is noticed as an image, however providing an alternative text is only optional for developers. Consequently, an image is not fully accessible, if it is correctly recognized as an image but does not offer an alternative text. In addition to that, even if alternative texts exist, there is no guarantee that they are helpful. Bad practices are when they only contain the file name of an image or when they start with words such as "image of ..." (Firth, 2024). According to Firth (2024), alternative texts need to be usefully expressed. Recommendations are to write them specific with a maximum length of 140 characters and without advocating the opinion of the writer. Furthermore, a text shown in an image must be provided in the alternative text, punctuations must be included and they have to be optimized for search engines (SEO). Referring to this, Levc says:

> *"[...], However, when it comes to image descriptions or alternative texts that are actively created, it is often up to the person creating the alternative text to decide what they think is important. [...]"* (Levc, 2024)

How accessible websites and images actually are with screen readers depends not only on the technical realization of the website, but also on the various capabilities of the different screen readers. This will be discussed in more detail in the next section.

2.1.2.4. Assistive Technologies

Blind people frequently use assistive technologies to overcome hurdles in their daily lives. These technologies can be considered as primarily technological tools coming from interdisciplinary research fields. These aids can exist in various forms such as technologies, equipment, environmental modifications or services. They provide support in cases like mobility, navigation, object recognition or social interactions, and are intended to ensure independence so blind people are equal participants in society (Bhowmick & Hazarika, 2017).

When it comes to the independence on websites, as previously discussed, blind people frequently use screen readers, that read out loud the content of a website. Therefore, they either choose the audio output or use the provided interface for other assistive technologies such as braille displays (Firth, 2024).

Levc (2024) reports that she uses a braille display as an input and output device in the office as well as at home on her computer. On her smartphone, she uses the VoiceOver function, which reads content out loud, to navigate, write and read text messages and much more. In addition to the standard functions, she also uses apps for blind people such as "SeeingAI" or "Be My Eyes". With these apps, a scene can be photographed and then be described. These apps are also connected to a community, which means that human assistance is offered if required.

The market offers a series of different screen reader products that have several features and run on various operating systems. Some of them are free and others are charged. The four most common screen readers are listed below:

- **Non-Visual Desktop Access (NVDA)** is a free software running on Windows that enables blind users to make many desktop applications (e.g. Microsoft Office or web browsers) more accessible. It also offers the support for over 55 languages and provides interfaces for braille displays and keyboards (NV Access, n.d.).

- **Job Access With Speech (JAWS)** runs on Windows as well and supports different software like web browser, Microsoft office, remote desktop and also PDF with optional braille devices. Special features are the smart picture recognition and the support for accessing printed media by using the camera. JAWS is a paid software. The price has a range from $95 per year to $1625 as one-time payment. Consequently, each rate has another offer of features (FreedomScientific, n.d.).
- **Apple VoiceOver** is the built-in screen reader from MacOS that can be used for several applications such as web browsers (Apple Inc., n.d-b). It provides two modi for navigation on websites: the navigation by DOM and the navigation by grouped objects (Apple Inc., n.d-a). Similar to NVDA all punctuations, text formattings and spelling errors are highlighted during usage (Apple Inc., n.d-b). As Apple VoiceOver is integrated in the operating system it comes without any costs.
- **Orca**, that is free and open source, has the capability to read out loud texts. It also provides users with an interface for braille devices (GNOME, n.d.-a). Furthermore, Orca offers features to navigate through websites, capture text formatting, fill out forms, interact with dynamic web content and also provides features for general orientation on the screen (GNOME, n.d.-b).

Jones (2018) presents a method that allows sighted people to simulate the user experience with screen readers, giving them a better impression of its handling. This experience is illustrated in Figure 2.3, where a piece of paper in the same a size as the screen is prepared and put on the monitor. This sheet needs a cut-out in the middle, that is large enough that two to three words of a sentence are fully visible. This simulation represents what blind people can perceive while navigating on websites with screen readers.

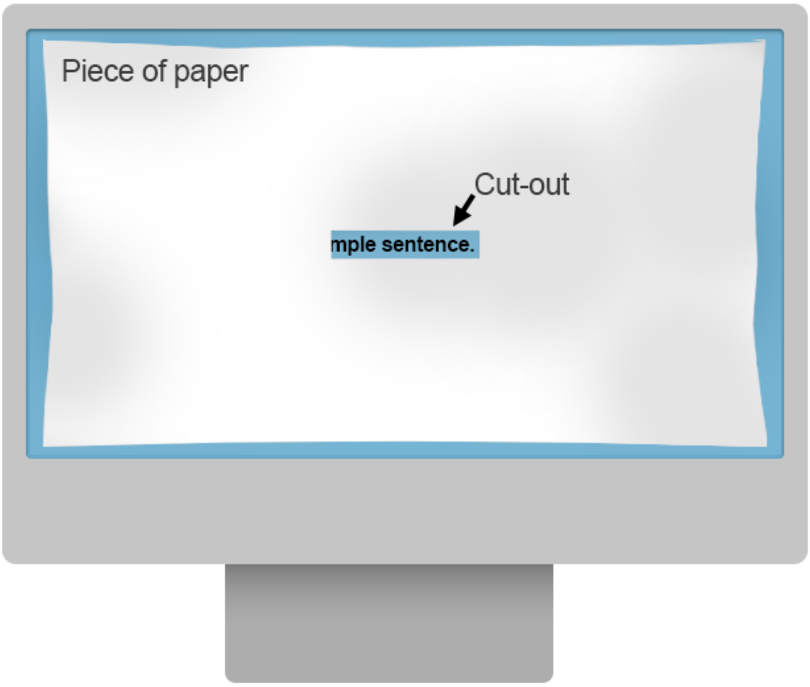

Figure 2.3.: Simulation for sighted people to experience a screen reader.

After having discussed the most commonly used screen readers, how they work and where both their advantages and disadvantages lie, the user experiences on websites of sighted and blind people can be compared.

2.1.3. How People Experience Websites

Most developers have little to no knowledge how to design a website for blind people, because they have no understanding how they navigate on websites. Oyewole (2019) describes this problem in the following statement:

> *"How do you go about designing or developing for these people when you don't even know how they navigate the web"* (Oyewole, 2019)

In order to get a better insight into the web navigation of blind people this section elaborates how blind people use websites from a practical standpoint and especially how they perceive images. Since websites are primarily designed for sighted people (Oyewole, 2019), the comparison of the user experiences of blind and sighted people should assist in identifying barriers. Therefore, the navigation behavior of both as well as their experiences with images will be discussed. To begin with, common navigation strategies by sighted users will be presented in the next section.

2.1.3.1. Sighted People Navigating Websites

A common method for analyzing the navigation behavior of sighted people using websites is eyetracking. Eyetracking is a method, where the trail of where a person is looking on the screen is traced. Therefore, the computer monitor has to be appropriately equipped and an eyetracking software can then be used to track user's focus paths on the monitor (Nielsen & Pernice, 2010).

Nielsen and Pernice (2010) applied this method to figure out different navigation patterns of users on websites. In this context, they started to research where sighted users look first when entering a website. As a result, there is no shared pattern where they look at, because people are individual. Nevertheless, two essential parameters influence where people might look at and these: firstly, the task they are following and secondly, the prior experiences from websites they visited in the past.

To be more specific, it is probable that people are first attracted by a flashy popup in the middle of the screen such as a cookie banner or an advertisement. Otherwise, users may immediately check the website's logo for verifying on which page they currently are. They could also draw their attention to the navigation panel or look at one of the headlines straight away (Nielsen & Pernice, 2010).

The example in the next Figure 2.4, which is adapted from Nielsen and Pernice (2010), shows how users navigate through web shops. In this case, the task was to find any item to buy, while an important factor of this research scenario was that users had to orient themselves on the website first. The adapted Figure 2.4 illustrates, that the user decided to directly look at the main menu, where confusion followed (this can be seen at the zigzag going across the menu buttons) and then the look went to the website's logo to confirm that the website is still the right one.

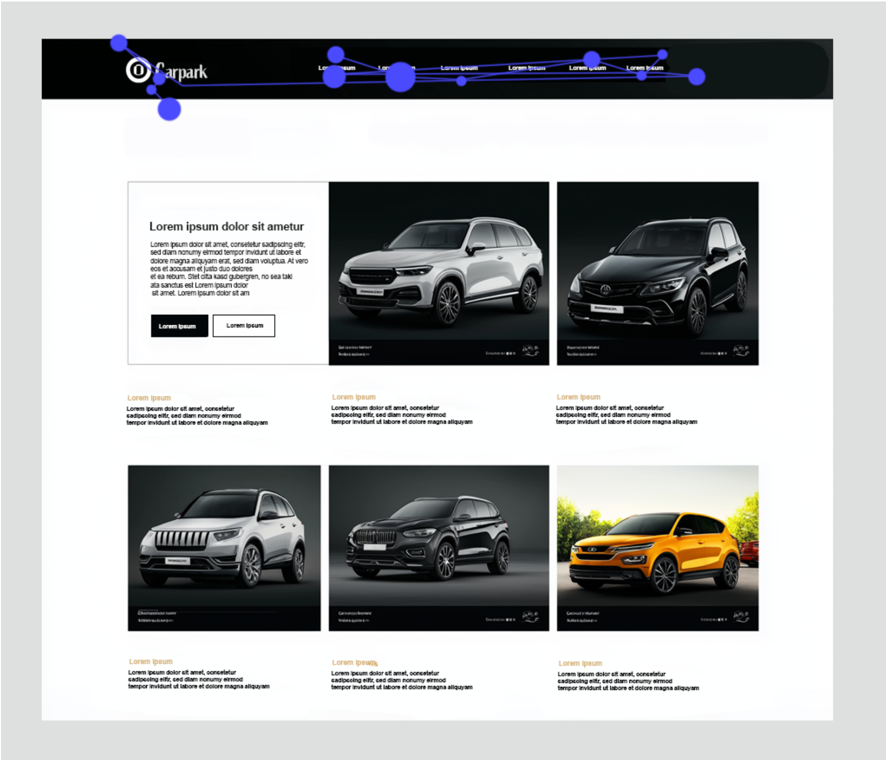

Figure 2.4.: Readjusted example of eyetracking in online shop.

As seen in this example, the navigation menu seems to be a place where users look at when they are trying to find a certain piece of information on the website. It might even be, that over a quarter of the time spent on a website is used for looking at the navigation (Nielsen & Pernice, 2010). According to Nielsen and Pernice (2010), this behavior is caused by users making a type of "safety check", so they feel comfortable when they are sure that the navigation menu is and stays there.

Another example demonstrated, how links and headings are used during navigation. In this part of the study, users were asked to find something they are interested in to read on a newspaper. In order to fulfill this task, one the one hand, they show a tendency to instantly look at sections' headings (e.g. "Latest News") and subsequently quickly scan through the corresponding links. On the other hand, they rather draw their attention to a features story that is shown on the home page (Nielsen & Pernice, 2010).

Links and headings play an important role in trust and efficiency for finding information. In fact, users often look at links and headings first in order find out, based on the displayed names of these elements, which sections of content are relevant to them and which they can directly skip (to have less effort) (Nielsen & Pernice, 2010).

However, according to Moran (2017), if headings and similar elements are not visible on a webpage, people tend to imitate the same behavior on the main text instead. To do this, users only scan the first few words of paragraphs or sentences, but even this attention span decreases the longer users are on the webpage. Since this navigation path approximately resembles an "F", this approach is defined as the F-Pattern. This method of scanning webpages is considered as efficient, but not effective (Pernice, 2019). This means that using the F-Pattern is fast, but important information can be missed.

Pernice (2019) presented a last pattern that is considered as the most effective one. In contrast to the previous patterns, in this one users read the texts on webpages carefully. This means, that they draw more attention to text passages and show a longer attention span. People might use this navigation method when they have built trust in a certain website, based on their past experiences, so they expect to find desired information on it.

Even though it might be impossible to exhibit all examples of how sighted people can navigate a website, the presented examples provided an insight on some commonly used efficient and effective methods. In order to better understand how blind people navigate the web the next section discusses this topic in detail.

2.1.3.2. Blind People Navigating Websites

Blind people use the internet with the aid of screen readers, because they compensate their vision loss with them. As mentioned in Section 2.1.2.4 the screen reader is a type of assistive technology and some of them even offer an interface to connect further assistive technologies such as braille displays or else (Sulong & Yusof, 2024).

As a fact, blind people are conquering the textual content with screen readers very fast. In many cases, their listening abilities are so well-trained, that they can listen to the screen reader in a pace which sighted people can impossibly follow. Depending on the situation, the pace of a screen reader can be adjusted at any time. This means, it is possible to switch between faster or slower reading (Oyewole, 2019).

Blind users also strive to find relevant content on websites as efficiently as possible. Instead of using a mouse, which would require sight, they use the keyboard for navigation (Oyewole, 2019). Scanning a webpage begins with finding orientation on a website first. Therefore, screen readers provide various features that can be used based on individual needs (Sulong & Yusof, 2024). As an example, Levc (2024) describes how she uses her screen reader to gain an overview on webpages:

> *"I always start by using the keyboard shortcut to go to where the screen reader defines "at the top", so to speak, and then I use the arrow keys to go down line by line to get an idea of how much menu there is before I get to the actual content and so on. And how barrier-free is the whole thing in general? This is the first exploration, so to speak"* (Levc, 2024)

According to Sulong and Yusof (2024) users reach individual decisions how to learn a website. On the one hand, this involves jumping from heading to heading or link to link (Sulong & Yusof, 2024). Levc (2024) adds that her screen reader is able to navigate only through headings by pressing the "H" key on the keyboard and she considers this method as very efficient. In particular, traversing the headings is often used to explore an unknown website, to read the web information architecture or to navigate in case the skip link does not work (Sulong & Yusof, 2024).

On the other hand, another frequently used way of navigating and finding relevant content on webpages is to use landmarks as e.g. "Home", "About", "Contact" and so on. These are used especially if the design of a website is complex and they want to find a specific passage. A further solution to navigate over complex layouts is to switch to the text version of a webpage, because navigating on it has a reduced complexity (Sulong & Yusof, 2024).

The search function is also commonly used for efficiency reasons. Some blind people always use this feature to navigate while others use it if they became lost on their paths, so they can return to the start. Also, the search function is used if shortcut combinations do not work and it is necessary to proceed differently. In addition to that, this function has a great benefit when it comes to reading large documents in a shorter amount of time, because users can easily skip parts of the content to get to specific sections (Sulong & Yusof, 2024).

Regarding text content on webpages, another asked question in the interview was how blind users cope with typos or spelling mistakes. Levc (2024) reports her experiences in the following quote:

> *"[...] I am a very accurate typo finder. The only thing it doesn't read really well is capitalization errors, but everything else. I don't miss anything else. Even with punctuation errors, i.e. the intonation goes with the emphasis and then you realize that something is wrong when the emphasis suddenly no longer fits. [...]"* (Levc, 2024)

She also indicates, that screen readers sometimes read out in the wrong language. This problem happens sometimes, because the language setting in the code, which defined during development, does not match the language of the actual provided text content (Levc, 2024).

Moreover, further hurdles are inaccessible interactive elements, especially concerning buttons and links. According to Levc (2024), if links are not properly defined as links in the DOM tree, they will only be announced as "link" by the screen reader without any further information. Also buttons can lead to problems when they are either not labelled properly (e.g. they only contain an icon) or their state (e.g. disabled) cannot be accessed by the screen reader. Hence, buttons can be both misleading and unclickable and this causes a challenge for blind users to be figured out.

2.1.3.3. Comparison of Navigation Behaviors

Since the most relevant navigation techniques of sighted and blind people were covered in the last two sections, a comparison of the findings will be presented. Table 2.1 is intended to show the similarities and differences in their navigation patterns.

Navigation	Sighted users	Blind users
Using assistive technologies		✓
Using keyboard navigation	✓	✓
Using mouse navigation	✓	✓
Scanning content	✓	✓
Navigation by headings	✓	✓
Scanning large documents	✓	✓
Navigation by search function	✓	✓
Using landmarks for navigation	✓	✓
Reading content accurately	✓	✓
Rely on accessible DOM		✓
Require human assistance for specific errors		✓
Easy handling of popups	✓	✓
Easy handling of faulty language settings	✓	✓

Table 2.1.: Comparison of navigation patterns on websites of sighted and blind users.

As a result (see Table 2.1) there is a relatively large overlap in the navigation patterns of sighted and blind people. The difference for blind people, which can lead to problems, is that websites are commonly neither properly designed nor developed for screen readers. If a website is designed for screen readers though, thus blind people can use them without complications. These complications often occur due to accessibility measures that are not taken into account during development. In this sense, sighted people have the advantage of being able to compensate for technical deficiencies thanks to their visual abilities.

2.1.3.4. Sighted People Accessing Images

As mentioned, the looks of sighted users often move between texts and images. This highlights the importance of images for the user experience on webpages. According to Flaherty (2017) companies are often using this kind of layouts when they are explaining sophisticated content, so images besides texts make it easier to follow the topic. In addition to that, it is common as well that this pattern is used repetitively throughout the content of a single webpage.

Moreover, websites can offer four modalities to communicate with the user. These are texts, images, videos and audio. Among these four, images are the most notable ones, because users usually recognize them directly (Nielsen & Pernice, 2010).

While users are scanning a webpage, they perceive images already at the edge of their visual field. This very short moment makes them immediately reach a decision, if an image seems worthy to draw their attention to it. This moment typically occurs in a time span within less than 20 milliseconds. Reflecting this behavior, it is assumed that users have a kind of "limited number" of glances that they can utilize on a website. Hence, decisions are made rapidly and every glance that turns out to be irrelevant influences how the webpage is subsequently viewed. As a result, only 42% of all images on a webpage are actually observed (Nielsen & Pernice, 2010).

Since images are obviously only viewed for a very short time, the question arises why they even exist on websites. The answer is, images are an important design element for websites as they can enhance certain feelings, provide information and improve the entire user experience. This implies however, that they have to be used in an appropriate way. Therefore, the developers have to choose the right images for a specific content, because well selected images can convey emotions and deliver messages much faster than texts (Nielsen & Pernice, 2010).

As stated above, images can significantly improve the user experience, however there are certain criteria involved that help images to be better recognized and to attract more attention (Nielsen and Pernice, 2010). According to Nielsen and Pernice (2010) these criteria are:

- Images must be colorful and produced with a high quality.
- Instead of shrinking images to a smaller size, they must be cropped.
- Interpretations must be straight-forward.
- Images must fit the context of the webpage.

In addition to that, certain types of content shown on images can make users feel drawn to images even more. This involves, inter alia, people with happy facial expressions, people that are oriented or looking directly at the camera, people that have highly attractive anatomies, delicious food that is possibly elegantly served and information that is brief and precise enough to deliver the meaning quickly (Nielsen & Pernice, 2010).

In contrast, images that are inappropriately designed are often avoided by users. Users want to be efficient, so they tend to evade these images, because these images increase their mental load while scanning a webpage. As an example, poorly designed images on webpages can tend to look like an advertisement. They can also be unrelated to the content or boring at the first glance (Nielsen & Pernice, 2010).

2.1.3.5. Blind People Accessing Images

This section is of particular interest for this thesis, because for contextually analyzing images it is essential to understand what blind users prefer to know about images. This is a great challenge since screen readers, as already explained, are only capable of reading textual information. Blind people gather their information from alternative texts (Tiwari, 2023), thus these texts must be written in such a way that blind users can benefit from them. According to Levc (2024) the way how alternative texts are written strongly depends on the person's writing and she claims that automatically generated ones often do not work.

> *"That depends on the person who creates the alternative text. I don't think there are any real guidelines yet. Automatic alternative texts often don't work at all [...]. But with the image descriptions or with the alternative texts, which are really created actively, so to speak, it is often really just the decision of the person creating the alternative text, based on what they think is important. So if, for example, there are pictures of some events on the university homepage in the news articles, then sometimes it is listed who is in the picture from left to right and just the people and their names, or a picture that just contains some object or something else is described. So it really varies a lot and depends on the picture and the person who is authoring it"* (Levc, 2024)

In the following, various concerns of blind users regarding which information from images is interesting in respect to the context of the images and issues of the websites are discussed. The study Stangl et al. (2020) describes many different scenarios, however only a few examples will be demonstrated.

When reading news paper articles blind participants state that it is important for them to know the reason for picking a certain image for a specific article. The exact required information strongly depends on the context of the article, e.g. if the focus of the image is a famous person, then the image description should explain certain characteristics about the person's identity. In other scenarios, for instance a sports team, the preference of blind users is to understand the interactions and the actions in general (Stangl et al., 2020).

The interviewees mentioned that Facebook is the best platform in terms of image descriptions. On Facebook they can grasp the name of the people involved and something about their actions and interactions in the picture. As good as it sounds, this is not the standard case, because there are many other examples from Facebook where the image descriptions are very poor and hence unusable. Generally, they prefer to know the facial expressions or body language of people on images, however if they know a person they show great interest for more specific details and information of the environment (Stangl et al., 2020).

According to Stangl et al. (2020) the blind participants claim they have never used online dating platforms. The reason is that they are totally inaccessible. Despite of that circumstances, they describe what they prefer to know about people they are dating. A very important aspect are physical characteristics of the person, e.g. body weight, body type, the hair or beard, eye color and also if the person has any special features like tattoos or moles. Additionally, they are interested in person's hobbies or passions, so they can try to understand the environment shown on the image. This can give insights to their lifestyles.

Levc (2024) mentions that the kind of information she expects from an image depends on the context. As an example, when she accesses an image that shows a landscape, thus she appreciates a detailed description of the scenery. In contrast, when she visits an online shop for fashion she expects a clear description of pieces of clothes shown (e.g. a dress). However, in case she accesses an image about an honorary doctorate award then the exact styles of the clothes are not relevant, instead the situation and the interaction of the people are interesting. Importantly, when a diagram is shown her personal preference is to grasp the results of it. Information as e.g. the colors of the diagram are irrelevant at this point.

2.1.3.6. Comparison of Image Perception

In this section, the previously discussed approaches of how blind and sighted people perceive images on websites are compared. To do so, the following Table 2.2 compares the results of Sections 2.1.3.4 and 2.1.3.5:

Images UX	Sighted Users	Blind Users
Using assistive technologies		✓
Perceive images visually	✓	
Perceive images without alt texts	✓	
Perceive images with alt texts		✓
Attraction by color, image quality and objects	✓	
Proper quality of alt texts required		✓
Individual interests for details of images	✓	✓

Table 2.2.: Comparison of the perception experiences of images on websites of sighted and blind users.

As a result, blind people have significantly fewer options for perceiving images on websites than sighted users. Even though the alternative texts can easily be accessed via screen readers, the actual quality of image descriptions strongly depends on their authors.

In conclusion, it can be said that blind people benefit from images just as much as sighted people. However, blind people often experience significantly more hurdles when it comes to freely deciding which details of images they want to perceive. With regard to general image information, it was shown that the context in which the image is found is crucial for defining which image details are relevant. However, when it comes to more in-depth details that blind people want to know about out of personal interest, they are very dependent on the author of the alternative text to include the desired information in the image description. In some cases, blind people cannot get any information from images at all because there is no image description available. On top of that, sighted people have a big advantage when discovering images on the webpage, because they do not have to explicitly navigate to the images like screen reader users do.

Since developers are often writing unsuitable or no alternative texts at all, Stangl et al. (2020) states that there are already tools to auto-generate image descriptions independently. However, these tools are partially successful, so the most recent Artificial Intelligence technologies could have the necessary potential to finally reduce these barriers. In this sense, various features of AI are presented in the next section, that may provide the necessary technological capabilities to provide blind people with context-sensitive image descriptions.

2.2. Artificial Intelligence

Although Artificial Intelligence (AI) has only been trendy for a short time, the concept has existed for more than 70 years. Nowadays, evolving computing power, massive amounts of data, proper algorithms and financial investments enabled AI to finally show its potential (Fumo, 2017). However, it is hard to precisely define AI, because the understanding of it has a wide range. There are numerous approaches in defining specifications that qualify a technology as "intelligent". Even though some definitions of AI give a good impression, they still have their limitations (Sheikh et al., 2023). A few approaches to define AI are shown below:

According to Sheikh et al. (2023) Algorithms are a widely used way to explain the function of AI. By definition, algorithms are a determined order of instructions for solving a problem. The critique of this understanding is that simple devices such as a pocket calculator would be considered as Artificial Intelligence too.

Another definition of AI is that computers are able to imitate human intelligence. Although this concept appears to be more sophisticated, if human intelligence is used as the criterion, thus present AI technologies would not meet the requirements for this version of intelligence (Sheikh et al., 2023).

Moreover, a more precise approach to define AI is to consider it as a technology that is capable of perceiving, taking actions and achieving goals in its environment. These performances should be executed with a certain degree of autonomy. However, the critique in this case is that the exact level of independence is hard to declare (Sheikh et al., 2023).

In summary, defining AI is a great challenge, since scientists have not managed to completely understand the mechanics of human intelligence yet (Sheikh et al., 2023), which makes it even harder to replicate intelligence artificially (Kulkarni et al., 2023).

This thesis follows the definition of Kulkarni et al. (2023) which states that AI refers to a machine that is executing tasks that would have typically required human intelligence. In order to achieve this, the machine has to learn "intelligent" behavior from training data.

Therefore, this section debates essential information about how machines are learning and applying human languages, in which ways generative models differ from conventional language processing, which role current language models play and how these areas are thematically linked. In addition to that, this is followed by strategies to improve prompts and general concerns of businesses and private users about Generative AI.

2.2.1. Natural Language Processing

It is becoming evermore important for companies to analyze text data that are mostly written by humans, because they contain crucial information. These data can be found on social media platforms, articles, websites etc. and are considered as unstructured data. Programming computers to understand unstructured data raises a challenge, because their capability is analyzing numerical data instead. The reason why computers have challenges to analzse text data is, that these data do not fit into any predefined data categories. Therefore, the discipline Natural Language Processing (NLP) provides many approaches to deal with this sort of data (Sarkar, 2019).

Before delving into the computational part of NLP, it is necessary to get an understanding what natural language (human language) is, how it is used and how it can be learned by humans. Firstly, Sarkar (2019) describes natural languages as languages such as German, English, Chinese and so on, that were primarily developed by humans while communicating with each other, rather than by explicitly defining rules. Secondly, the purpose of languages is to deliver a meaningful message to other people. In order to do that, words have to be concatenated to sentences (Bonvillain, 2020) by correctly applying syntax, semantics and grammar (Sarkar, 2019). Words themselves are used for representing e.g. people, attitudes, situations, thoughts, objects etc., and sentences are used to describe situations in different contexts (Bonvillain, 2020). Finally, acquiring a language is a process that goes over many years. It starts with children who are exposed to surroundings, where they have to hear a certain language all the time and need to actively use it for interacting with others (Al-Harbi, 2019). In the course of this, their language skills enhance by imitating from sentences they heard and by extracting patterns out of it (Sarkar, 2019).

Consequently, Natural Language Processing (NLP) is a way for humans to interact with machines by using human language. This is an evolving discipline coming from AI and Computational Linguistics (CL). Its methods enable machines to interpret text data, so they can understand their contexts and even properly respond to them. Nowadays, machines offer enough computing power for running neural networks. Researchers utilize neural networks for conducting text data analysis, instead of using probabilistic methods. Therefore, neural networks are trained with a large amount of text data, where the machines are focused on finding patterns and relationships (similar to humans in their language acquisition process), so they can develop the capabilities to predict the next word in a sentence depending on its preceding word (Kulkarni et al., 2023).

As this has led to enormous advancements in developing chatbots, language translation programs and so on, it appears promising to revolutionize the overall way of human-computer-interaction in the future (Kulkarni et al., 2023).

2.2.2. Generative AI

As mentioned, via NLP methods, machines can learn to understand text data and predict e.g. further words in a sentence. This refers to discriminative models that are trained to categorize data (e.g. by labeling them) and use them for making forecasts (Foster, 2023).

In contrast, Generative AI is able to generate brand new data instead of making predictions (Foster, 2023). This involves data types such as texts, images, videos or audio, however the generation of text has the greatest impact, because it finds the most applications in nowadays businesses (Kulkarni et al., 2023). In order to produce new content, the computer has to be trained with training data to develop a generative model. Figure 2.5 gives an overview how this process generally works. Therefore, it is essential that the generative model gets trained by observing training data that resemble the desired output (Foster, 2023). The desired output is described by a prompt, that is basically a text message written by the user, to instruct the AI to execute the generation process (El Amri, 2024). As a result, by additionally using a random noise the AI is capable of generating new content, that has not existed in the training data before (Foster, 2023).

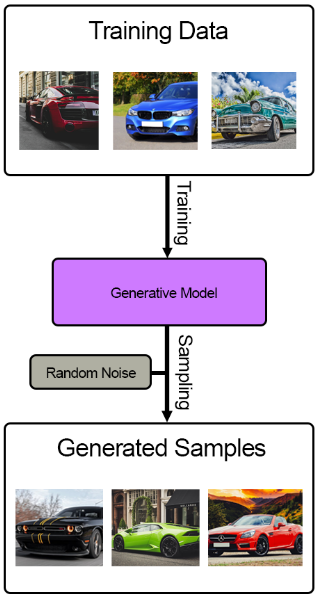

Figure 2.5.: Process for generating new content from existing data. Based on Foster (2023).

Figure 2.6 shows where Generative AI can be classified among other AI disciplines (Kulkarni et al., 2023). According to Kulkarni et al. (2023) Generative AI is a specialization of Deep Learning (DL) that is trained to create new content. Firstly, Machine Learning (ML) uses particular algorithms for enabling the computer to acquire "intelligent" behavioral patterns, whereas secondly, Deep Learning is a more in-depth form of ML, that uses neural networks to analyze complex coherence of data in detail. As a result, Generative AI is a form of DL that is fine-tuned with the purpose to make new content (Kulkarni et al., 2023).

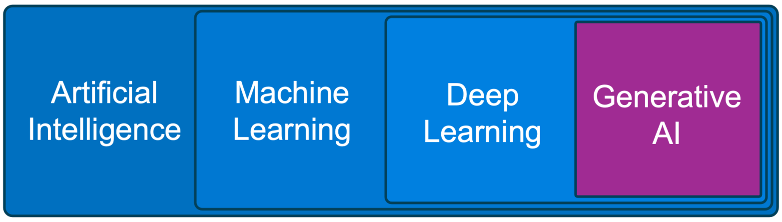

Figure 2.6.: Interplay of artificial intelligence, machine learning, deep learning and generative AI. Based on Kulkarni et al. (2023).

Since generative models in the field of AI were thematically classified in this section, the next section deals, on the one hand, with different generative models and, on the other hand, with the connection between Generative AI and NLP.

2.2.3. Large Language Models

A Large Language Model (LLM) is a model that enables the machine to understand and generate natural language. In order to learn that, LLMs need to undergo certain trainings with massive amounts of text data. Therefore, Natural Language Processing offers a variety of methods for language training and language data analysis (Ozdemir, 2023). As a result, Generative AI uses the capabilities of Natural Language Processing that is driven by Large Language Models.

However, the mentioned text data for training are considered as unstructured data (see Section 2.2.1). In order to use them with NLP methods, they must be converted into a better structured format. Therefore, whole sentences are broken down into sequences of small fragments. This process is called Tokenization and individual fragments are called tokens (Kulkarni et al., 2023). The following example should give a general impression of what Tokenization is doing:

"My pet is a cat." can be broken down into a sequence of tokens, namely "My", "pet", "is", "a", "cat", "." (Dop, 2020).

After that, sequences of tokens are processed through specifically trained neural networks, in order to analyze their context (meaning). In comparison to conventional ways, Generative AIs therefore utilize the so called Transformer architecture. However, it is not relevant for this thesis, to explain this architecture in detail, so only a very general insight will be given to explain the principle. The central component of Transformers is the so called Attention Head (Foster, 2023). Foster (2023) demonstrates this mechanism with an incomplete sentence such as "The pink elephant tried to get into the car but it was too". Obviously, this sentence has to be ended with a word, but which one is not determined yet. In order to figure out the ending word, attention must be paid to the context of the entire available part of the sentence. Therefore, some relevant words are picked out by the algorithm and irrelevant words, i.e. conjunctions or articles, are ignored. Finally, in this case a word like "big" (or a similar one) would be suitable for ending this sentence. A further reason to use Transformers is, that they can easily be parallelized which enables them to work fast with large amounts of data. Therefore, multiple attention heads can be used simultaneously, which refers to Multihead Attention.

Currently, there are dozens of different LLMs on the market, thus it is not possible to compare all of them within the scope of this thesis. This is the reason why LLMs from two companies, with different occasions, are selected for this research. Firstly, the GPT-family from OpenAI is chosen, because they were the first company that released an AI technology which is accessible and understandable for the majority of people (Schwedux, 2023) and secondly, the opponent is a well-known open source LLM-family published by Meta, called LLaMA (Carolan et al., 2024).

It should be noted, that OpenAI offers two types of services for using their models: the web version ChatGPT[7] as the easy-to-use version for the society and the OpenAI Application Programming Interface (API)[8] that is an interface for developers to use the GPT-family through their own software (El Amri, 2024). OpenAI offers the main GPT-models for both ChatGPT and the API, however the API encompasses additional subversions and has different usage conditions. The tokens for the API are charged per million, so they are officially labeled as the total cost for one million used tokens for input and output (OpenAI, n.d.-a). The Table 2.3 shows an overview about the LLMs of OpenAI (excluding the subversions), its available features, performances and costs (OpenAI, n.d.-b).

[7] **https://chatgpt.com/**.
[8] **https://openai.com/index/openai-api/**.

Model	Capabilities	Context Size	ChatGPT	API[a]
GPT-3.5	simple tasks text input/text output audio input/audio output	16k	free	-
GPT-3.5 Turbo	simple tasks text input/text output	~16k	-	\$0.50/\$1.50
GPT-4	high intelligence text & image input/ text output	128k	free	-
GPT-4	high intelligence text & image input / text & image output	128k	paid	-
GPT-4 Turbo	high intelligence text & image input / text output	128k	-	\$10/\$30
GPT-4o	high intell. & fastest text & image input / text output	128k	free	-
GPT-4o	high intell. & fastest text & image input / text & image output / audio input / audio output	128k	paid	-
GPT-4o	high intell. & fastest text & image input / text output	128k	-	\$5/\$15

Table 2.3.: Specifications of various GPT versions.

[a] Dollar per million.

Meta's LLM series was published with the purpose of supporting researchers, that do not have the availability or access to a high-end infrastructure (Meta, 2023). It comes with the possibility to be served locally which offers a huge benefit, because it can be integrated in own applications and developers keep full control over it (Awan, 2024a). However, despite of not requiring a high-end infrastructure, the local computer still has to provide a certain amount of computational power (Witt, 2024). For that reason, even though the models are generally free to use, the expenses for necessary hardware have to be taken into account. Table 2.4 gives an overview about the LLMs and Large Multimodal Models (LMM) from Meta:

Model	Capabilities	Context Size	Costs	Parameters
LLaMA-1	text input / text output	2k	-	7B, 13B, 34B, 70B
LLaMA-2	text input / text output	4k	-	7B, 13B, 34B, 70B
LLaMA-3	text input / text output	8k	-	8B, 70B
LLaVa	text input & image input / text output	-	-	7B, 13B, 34B

Table 2.4.: Specifications of various LLMs and LMMs by Meta. Information are taken from Touvron et al. (2023), Awan (2024b) and Ollama (2024).

Apart from understanding how Large Language Models work, it is also important to know how to use them correctly. Guidelines on that are discussed in the next section.

2.2.4. Prompt Engineering

Prompt engineering is a discipline that focuses on ensuring that Generative AIs achieve better results that are closer to user expectations. Therefore, it is necessary to know certain tricks and strategies (El Amri, 2024).

These tricks and strategies are divided in two aspects: Firstly, prompts can be written in a properly expressed and structured way, and secondly, LLMs provide parameters that can be adjusted by the user so that the generated results are more aligned with expectations (El Amri, 2024).

To begin with the first example, Listing 2.1 El Amri (2024) demonstrates a certain sentence structure how users tend to express prompts.

```
USER: May I ask you to make a list of top 10 sights in New York City?
```

Listing 2.1: Prompt starting with a phrase. Example based on El Amri (2024).

This prompt can be improved by the trick of starting sentences with action verbs, rather than with expressions such as "May I ask ...", as shown in Listing 2.2. This makes the prompt easier for the computer to understand (El Amri, 2024).

```
USER: Make a list of top 10 sights in New York City.
```

Listing 2.2: Prompt starting with an action verb. Example based on El Amri (2024).

Another example is that providing more contextually relevant details helps to get more precise results. The Listing 2.3 demonstrates an inappropriate approach how a request should not be authored. In this case, the LLM does not get any necessary information about the context, so the AI is trying to guess its meaning (OpenAI, n.d.-c).

```
USER: 'Whos president?
```

Listing 2.3: Prompt shows a simple and open question. Example taken from OpenAI (n.d.-c).

In contrast, this is a demonstration of how OpenAI recommends to improve the previous Listing 2.3. The solution is shown in Listing 2.4 (OpenAI, n.d.-c).

```
USER: Who was the president of Mexico in 2021, and how frequently are
    elections held?
```

Listing 2.4: Prompt shows a question with details about the context. Example taken from OpenAI (n.d.-c).

El Amri, 2024, states an example that can be especially useful for tasks like coding. It is possible to use delimiters in the prompt, such as quotation marks, for separating different sections of the text by their purpose. The Listing below (see 2.5) presents this strategy, where the user queries the AI in human language to fulfill a task with the HTML code that is provided in the quotes underneath.

```
USER: Take the code in quotes and remove all CSS-classes.

"<div class='main-content primary'>This is an article about prompt
   engineering.</div>"
```

Listing 2.5: Prompt illustrates the usage of text separators for coding. Example based on El Amri, 2024

Both sources, El Amri (2024) and OpenAI (n.d.-c), provide far more examples for formatting and structuring prompts properly. However, showing all examples would exceed the scope of this research, so for more information see the references.

As mentioned above in this section, prompt engineering does not only require prompts to be well structured, but it also works with adjusting arameters of LLMs. ChatGPT is an exception, because the web interface has these parameters predefined in order to guarantee simplicity for non-technical users (El Amri, 2024). These parameters, according to El Amri (2024), are described in Table 2.5:

Parameter	Description
Temperature	Determines the variability of generated output. The higher this value, the more random the response.
Top-p	It is a percentage value that indicates the least probability a word must have for being used in a response. As a result, the pool of possible words is narrowed down.
Top-k	Specifies the maximum size of the pool of possible next words for the response. The words are sorted by probability, starting with the most likely one. The higher the value, the more variable the result.
Sequence Length	Sets the exact number of tokens used for a response.
Presense Penalty	This value, which can be both positive and negative, indicates how strongly existing tokens in the response must not be reused. Negative indicates more token repetitions and positive indicates fewer token repetitions.
Frequency Penalty	Specifies the repetition rate of tokens. A positive value decreases the frequency of token repetitions overall, whereas a negative value increases repetitions. (Values go from -2.0 to 2.0)
Number of Responses	Defines the exact number of responses generated for a prompt. The default value is 1.
Best of	Selects the best generated response based one the highest log-probability per token.

Table 2.5.: Parameters for optimizing the responses of Large Language Models.

Overall, prompt engineering does not provide a "perfect" approach for using LLMs. The way of drawing it up strongly depends on the context and the desired generated results. Obviously, previously mentioned factors such as performance and cost should also be respected, since they have an impact on the outcome.

Generally, the fact that LLMs can achieve new and unique results through the use of suitable prompts and parameterization offers many possibilities. However, it should be noted that many mistakes can still be made when communicating with AI, for example, the AI trying to guess users' intentions when using a prompt. The following section discusses possible dangers and concerns in more detail.

2.2.5. Threats and Privacy Concerns

As already discussed, Generative AI is a special kind of AI that enables users to create brand new content (see Section 2.2.2). Many people use Generative AI with the intention of making their work significantly easier. This includes generating social media posts, campaign materials, summarizing emails or even creating business plans. All of these things can be done at breathtaking speed by executing instructions given in prompts, and there is no need to outsource the work to others (Caruana & Goater, 2024). Besides the fact that this is advantageous for many purposes, two types of resources are required to operate these tools, namely very high computational power and a training with a huge amount of data (see Section 2.2.3).

In particular, the use and collection of data raises security concerns, as these tools are in fact used by the global public. A few threats to private users may include biased information, social damage or the spreading of fake news (Derner & Batistič, 2023). Moreover, in the sense of companies of different scales, this way of data acquisition can lead to legal, financial and also ethical issues (Baxter & Schlesinger, 2023).

The core problem lies in the data collection, as this is absolutely necessary for Generative AI to achieve the desired results as precisely as possible. Therefore, users are required to provide as much details as possible in their instructions (Caruana & Goater, 2024). The next statement briefly explains the issue of data leaking with LLMs:

> *"[...] data is added to the LLM and then usable by others that access the same Generative AI tool. The data you feed into Generative AI about your question is also captured and stored in its memory and can be exposed to third parties."* (Caruana & Goater, 2024)

Due to the fact that the data is made publicly accessible through the use of these tools, criminals can access it using workarounds. According to Kosinski and Forrest (2024) an example for such a risk are prompt injection attacks. In this sort of attack, criminals utilize the possibilities of LLMs to issue harmful prompts that compromise the security of Generative AIs. Therefore, they write prompts that seem legitimate to the AI, although they contain malicious instructions. By doing so, they can be able to exploit e.g. sensitive personal data, infiltrate systems that run LLMs through software interfaces or spread manipulated information on campaigns.

When it comes to fake news or false information in general, one of the frequently overlooked main implications are AI hallucinations. These refer to generated content, that seems reliable at the first glance, but they significantly deviate from the training data (Siebert, 2024). Since the training data were obtained from many different sources that contain e.g. historical and cultural content, LLMs can also tend to generate racist and (gender) stereotypical patterns. As an example, this carries the risk that women may be categorized as less competent than men in generated data (O'Connor & Liu, 2024). The causes of such AI circumstances may lie among other things in the training methods used to create the LLMs or in the algorithms of the generative models (Siebert, 2024). This is also a symptom of the fact that Generative AI is not really transparent and therefore it is not exactly known how the content creation data is used (Nature, 2023).

These concerns represent only a small part of the problems that arise or can arise from Generative AIs. However, in the course of this research, it has become clear that factors such as data security can be controlled to a certain degree. Therefore, special attention should be paid to this in the course of the prototype development.

2.3. Summary

When people use websites they want to find relevant information as efficiently as possible. In most cases, websites are designed in a way that they are visually appealing, so they are suitable for sighted people. Consequently, this leads to barriers for blind users. In order to overcome these barriers, they use assistive technologies to compensate their lack of vision.

In particular, images, that typically appear on websites, create a massive barrier, these are extremely important elements though as they can significantly improve the overall user experience. Generally, images often appear in the context of text sections and have the ability to concisely and efficiently convey information to sighted users. They usually look at an image for just a fraction of a second, but in that time they can grasp the information that is of interest to them. Since blind people cannot visually perceive images, screen readers are used to read out loud image descriptions. When image descriptions are written in a detailed and precise way, accessibility is ensured. However, a tremendous number of websites exist where image descriptions have not been (adequately) considered when developing a website. As a result, blind people receive either very poor image descriptions or none at all and thus they are excluded from information.

A solution to face the problem with inappropriate image descriptions could be the application of AI technologies. These technologies were launched on the market during the recent years and have the potential to create image descriptions autonomously. Due to the context awareness of LLMs they have the strength to recognize and describe only relevant image details which are interesting for blind users. Nevertheless, it has to be taken into account that AI requires a large amount of data and very high computing power to reach the full potential. Furthermore, the AI tools might cause costs for the users and they come with a risk which can affect anyone. These are concerns about data security and disinformation.

The question if it is actually possible to use Generative AI to make images on websites more accessible is examined in the next chapters of this thesis. Therefore, the development of an AI prototype, that generates context based image descriptions, will be documented and evaluated.

3. Design & Implementation

In this chapter the development of a functional prototype that uses Generative AI to make images more accessible for blind people is documented. The results of this chapter are of particular importance for achieving the research objective.

First, the intended goal (see Section 3.1) and user target group (see Section3.2) for the product development are defined, along with the main features of the product from a theoretical point of view (see Sections 3.3 and 3.4). Next, in Section 3.5 the most relevant approaches and results of the empirical research for this software development are elaborated. This involves both the application of Generative AI and pure JavaScript methods. Subsequently, the most important implementation steps are presented in Section 3.6, that were necessary to create a prototype that can be used in practice as a proof of concept. Finally, a short demonstration of how the prototype behaves and whether it meets the defined requirements will be given in Section 3.7.

3.1. Starting Point and Motivation

Findings of the previous Chapter 2 state that blind people have efficient strategies to navigate on websites. Therefore, they use screen readers and eventually additional interfaces such as braille displays. Even though navigating works well in most cases, a specific downside is that images are often inaccessible, as described in Section 2.1.2.3. These circumstances often occur, because developers commonly implement images primarily for sighted people and do not take accessibility measures such as alternative texts with useful information into account. Consequently, blind people are excluded from the experience of perceiving images (see Section 2.1.3.6).

As a solution, the primary goal of this thesis is to develop a prototype, that autonomously produces image descriptions with Generative AI. Therefore, the prototype should only focus on image details, that are relevant to be described in the sense of the webpage's context. As a result, firstly, automatically describing images on demand should improve the information accessibility and make blind people more independent, and secondly, having only the contextually relevant information described should enhance their navigation efficiency, because they could scan a website faster. However, it is crucial that selecting relevant details from images does not cause them to lose important information.

The second goal of this thesis is to evaluate the specifications of Generative AI for implementing this prototype. This involves testing the multimedia analysis and generation capabilities of LLMs, estimating the costs for development as well as production, and figuring out the requirements to provide a proper user experience.

3.2. User Target Group

The target group for this research encompasses blind people. Characteristics such as age, gender, job etc. do not play a role. In order to access websites and especially images, blind people require screen readers. There are different types of screen readers available for desktop devices on various operating systems (see Section 2.1.2.4), thus it is important for the target group to have at least little experience using them on websites. They can use their preferred screen reading software with their own customized settings and they can also utilize any additional device such as braille displays, because this will not affect the outcome of this thesis.

3.3. Concept

The idea for developing this prototype is to use Generative AI for making images more accessible and enhancing the navigation efficiency on websites. This concept is divided into two chronological steps, where the first one is illustrated in the Figure 3.1 below. It shows a screenshot of a section of the Wikipedia article named "Visual Impairment"[1]. The image within the yellow frame presents the blind man with a guide dog walking in a shopping center. The red marked area exhibits a content passage that contextually refers to the image. It explains the purpose of a guide dog and how the blind person interacts with it.

[1]https://en.wikipedia.org/wiki/Visual_impairment

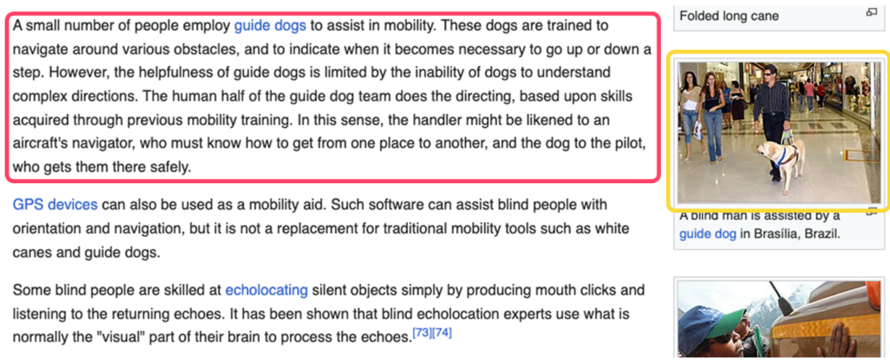

Figure 3.1.: Screenshot of a section of the Wikipedia article about Visual Impairment, the selected image is highlighted in yellow and the corresponding text is highlighted in red. Taken from User-duck (2024).

After the image in the yellow highlighted area has been classified as contextually relevant by the AI, it should be described in detail. In doing so, attention should be paid to include only those details in the description, that correspond to the text next to the image (see red area). In this case, an appropriate description could be worded as follows:

> *"The image shows a man, appearing to be visually impaired, walking through a shopping mall with the assistance of a guide dog. The dog, wearing a specialized harness, is focused on guiding the man safely through the busy public space, helping him navigate obstacles and maintain independence. The indoor setting, with various stores and people in the background, reflects an accessible environment where the guide dog plays a crucial role in supporting the man's mobility and confidence in navigating crowded areas."*

As a result, the following Figure 3.2 exhibits the Wikipedia article with the new image description appended at the end of the last sentence of the original paragraph (ending with "... who gets them there safely."). The yellow rectangle highlights the selected image, the original text is shown in the red rectangle and the newly appended text is marked with the blue rectangle.

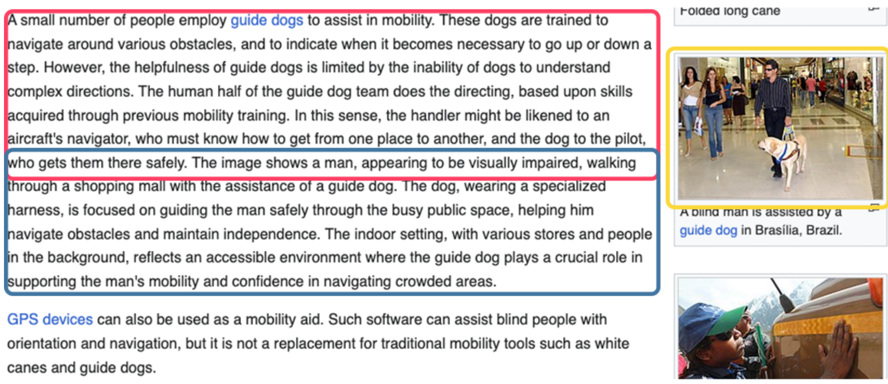

Figure 3.2.: Screenshot of a section of the Wikipedia article about Visual Impairment, the selected image is highlighted in yellow and the corresponding text is highlighted in red. The generated image description is highlighted with a blue rectangle. Taken from User-duck (2024).

Besides the context-based generated image descriptions, this prototype also has the goal to improve the navigation efficiency. Since each generated image description should be appended to the closest context-wise matching paragraph, there should be no further need for the user to explicitly navigate to the image, because the required image information are already in the text.

Additionally, a requirement is to ensure a user-friendly interaction based on the principles of usability from the Section 2.1.2.2. Since this prototype is designed to require only a small user interaction to start the image description process, all further processing steps should run independently of any user input. Usually, the generation of AI responses takes some time (e.g. 10 seconds), so a suitable user interface for screen readers should provide a status notification to inform about the progress of the image analysis.

3.4. Requirement Analysis

A substantial part of the prototype development consists of defining software requirements. According to Susheela et al. (2023) these can be categorized into two main categories, namely functional requirements and non-functional requirements:

- **Functional:** These requirements define what a software must be able to do and what user action is needed to accomplish that. In this regard, the main focus on this sort of requirements lies on the needs of the user. In case these requirements are met, the product is considered to be functionally complete.

- **Non-Functional:** The focus of non-functional requirements lies on how the software operates. This describes the traits or characteristics, such as the aesthetics, performance or availability of the product. In addition to that, these requirements have to be quantifiable, thus the success of the product can be measured by it. Nevertheless, these measures have almost no affect on the actual functionality of the software.

Having explained the basics of the requirements of both categories, the next two sections will elaborate on them in relation to the developed prototype.

3.4.1. Functional Requirements

Based on Susheela et al. (2023), the functional requirements of this project are explained. The elaborated list is presented below:

1. **User Action:** The entire image description process must be executable by pressing a keyboard shortcut.
2. **Algorithm:**
   a) **Data Selection:** Images and contextually related texts must be automatically detected on the webpage. To do this, the software must iterate through all images and text elements and select these, that are part of the main article's content. This step is essential, because certain elements that use images for their appearance are not considered as relevant in this case (e.g. icons or buttons) and consequently must be excluded from this selection.
   b) **Data Extraction:** The selected image-text-pairs must be extracted as JSON and they must keep references to their original positions in the HTML file. Hence, data at the original elements can be directly modified with new content if required.
   c) **Image Descriptions:** The extracted image-text-pairs must be each composed as a prompt that queries the Generative AI. This query has to select relevant details from the images, based on the provided text, and describe them. These image descriptions must have a certain length and be authored in the same language as the webpage.
   d) **Text Updating:** Once a new image description has been generated it must be appended to the original text on the webpage (by using the HTML reference from the data extraction). In order to improve the user experience by reducing the wait time for the users, the image descriptions must be appended one by one. Hence, the users can keep navigating on the webpage and view new image descriptions, once they have appeared.

3.4.2. Non-functional Requirements

The non-functional requirements (NFR) for this prototype research and development have to be defined, in order to evaluate its quality (Paradkar, 2017). Based on Paradkar (2017) the NFR specifications for this prototype are listed in the following:

- **Performance:** This parameter depends on two factors. Firstly, the total amount of images on a webpage, and secondly, the current state-of-the-art LLM that can be used. Apart from that, the AI should take between 5 to 10 seconds per image to generate a new image description.
- **Reliability:** For this prototype it is guaranteed to work on Wikipedia pages. However, if this software works successfully on other webpages strongly depends on their DOM-tree.
- **Maintainability:** The tool should be designed in a way, that adjustments and optimizations can easily be done, in order to apply changes regarding LLM developments in the future.
- **Compatibility:** the software must be able to run in Chrome Browsers on Windows, Macintosh or Linux.
- **Availability:** The product itself has no downtime, however its usage, in the course of this thesis, is limited to Wikipedia pages, since the DOM-tree is the most suitable.
- **Security:** Users must be aware of not using this program on webpages that contain personal or sensitive data.
- **Usability:** The program works autonomously, for this reason the user should only press a keyboard shortcut and the rest will be done by the algorithm. Success or failure should be announced in a way, that a screen reader can access it.

The functional and non-functional requirements are considered as a structured guideline, that enable the research and development of the product. The documentation of the empirical part is presented in the next section.

3.5. Feasibility Analysis

So far, a concept explanation has been given that describes the prototype to improve the accessibility of images on websites and the navigation efficiency for screen readers. From here on, an empirical and iterative research process was initiated, through which the prototype will be implemented. The focus of the development is on the practical application of Generative AI tools, which are evaluated extensively throughout the project.

The minimum requirement of this development is that the resulting prototype fulfills at least the functional requirements, stated in Section 3.4.1. Consequently, it has to be technically robust enough to be tested by the target group.

In the following section two approaches will be evaluated. Both approaches involve, on the one hand, a method for extracting images and context information from a webpage, and on the other hand, a tool for describing the images based on context details. The first approach addresses conducting the whole process with AI, while the second approach deals with a combination of a non-AI and AI implementation.

3.5.1. Approach 1: Webpage Analysis with AI

Approach one discusses the extent to which Generative AI can completely autonomously be used to extract images and corresponding texts from a website, without any manual programming intervention. Furthermore, these data should be processed by AI in such a way, that the desired information can be obtained and stored as JSON, in order to make a further development with these data possible.

As part of this approach many practical experiments were carried out. These involved several Large Language Models from both companies OpenAI and Meta. Specifically, for Meta's LLMs, the research was carried out on two platforms: firstly, on the local private MacBook[2], and secondly, on a Shadow PC[3]. When comparing the used prompts with each other, some of them only had slight changes (e.g. differences in wording or their formulations) while others were significantly different in their structure, content and length. In any case, a detailed explanation of all the tests would reach beyond the scope of this documentation, thus only the two most important attempts, using LLMs from OpenAI and Meta, will be documented.

The following Table (see 3.1) presents the technical specifications of the Shadow PC that was used in the course of this research. This remote computer was rented for two months, where the costs for each month were 74.99 € (Shadow, 2024).

[2]https://support.apple.com/en-us/111932
[3]https://shadow.tech/

Component	Specification
CPU	Intel Xeon™ (up to 4.5 Ghz) 12 vCores
GPU	NVIDIA® Quadro RTX 6000 24 GB GDDR6
Memory (RAM)	41 GB
Operating System	Windows 11

Table 3.1.: Hardware specifications of the rented Shadow PC (expert version). Information are taken from Shadow (2024) and NVIDIA (2019).

Importantly, the Generative AI is used mostly as a *Parser* in this scenario. This means that the AI is not only supposed to answer a prompt creatively, but data are provided and these must be processed one hundred percent correctly. In other words, if the AI decides to select e.g. a certain paragraph, thus it must be extracted without cutting off sentences or hallucinating content. This reliability is necessary for the project to be successful.

As mentioned above, a lot of prompts and generated responses were used in this empirical research. The OpenAI API was financially not suitable for this research, because it charges for each prompt and generated response. In comparison to that, ChatGPT comes with a fixed rate per month that will not increase. Hence, ChatGPT was used during the iterative testing process for prompts (see Section 2.2.3). OpenAI API, however, provides an interface for custom software to communicate with the LLMs, so this is used for the final product.

All the tests presented in this documentation were carried out using the English Wikipedia article of "Marie Curie"[4]. The choice of the Wikipedia article about Marie Curie was a coincidence —these experiments could have been conducted with other Wikipedia articles as well. However, it was found that Wikipedia articles were a proper choice for simplifying this research, since they mostly have no asynchronous changes in the HTML, and they met the necessary text requirements (containing images and text paragraphs).

The exact process of this research is illustrated in Figure 3.3. It can be seen that both inputs, so firstly the HTML file (A) is processed with the text analysis capabilities of ChatGPT-4o and LLaMA-3:70b, and then the screenshot (B) is analyzed with the vision capabilities of ChatGPT-4o and LLaVA-3:34b. All four attempts were targeted to generate a JSON file containing the images of the Wikipedia article and the corresponding images.

[4]https://en.wikipedia.org/wiki/Marie_Curie

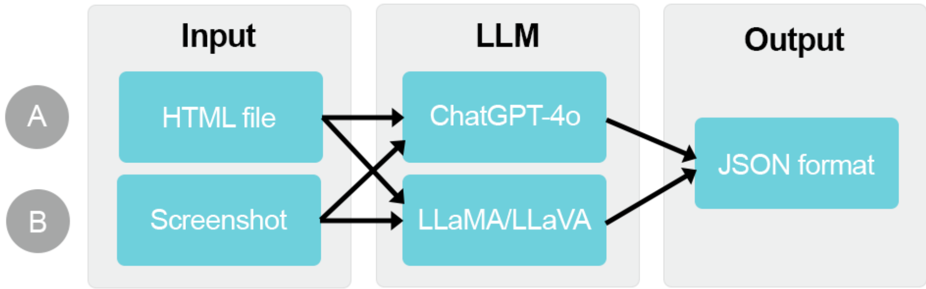

Figure 3.3.: Structure of the sections to compare the HTML file and screenshot analysis with ChatGPT-4o and LLaMA-3:70b or LLaVA-3:34b.

3.5.1.1. A: Analysis of HTML-File

The goal of the following experiment was to examine, whether ChatGPT-4o or LLaMA-3:70b can be provided with the HTML file of a webpage, in order to extract contextually related text and image pairs from it.

Both images and textual content can be implemented on a webpage in many different technical ways (see Section 2.1.2.3). Thus, a part of the challenge was to find out, if Generative AI, with its contextual understanding and programming abilities, would be able to correctly filter out relevant images and texts from an HTML file.

Therefore, the requirements for these attempts were:

1. Texts must be of complete length (without cut-offs).
2. Texts mut not contain any spelling or wording mistakes (no hallucinations).
3. Images must not be downloaded as files, instead their source URLs should be picked.
4. The filtered and extracted data must be collected in a JSON format, thus further processing in the development is easier.

In order to perform these tests, the Wikipedia article was downloaded as a file and attached to the prompt. The prompt, that was used for both ChatGPT-4o and LLaMA-3:70b, is shown in Listing 3.1 and was expressed using strategies from prompt engineering of Section 2.2.4.
It has to be taken into account, that ChatGPT-4o has the option to directly attach files in the GUI, while LLaMA-3:70b only takes the file's content as a string via the command line. The first line in Listing 3.1 tells the Generative AI to act as a web developer and provides meta information about the HTML file. Thereon, lines 3 to 5 represent a step-by-step guideline what the AI has to accomplish.

Finally, in the last three lines (line 7 to 9) of the prompt the desired output format for the generated response is shown (JSON).

```
1 You are a web developer. I provided you with a HTML file of the
      english wikipedia article about 'Marie Curie'.
2
3 Step 1 - Extract all texts based about Marie Curie and write them to
      the JSON in triple quotes. Ignore search bars, navigation, logos,
      agendas, login, signup.
4 Step 2 - Find all images referring to Marie Curie in write the
      imageURLs to the JSON in triple quotes. Ignore search bars,
      navigation, logos, agendas, login, signup.
5 Step 3 - Only output the JSON (pretty print) and don't output any
      other answer text.
6
7 """
8 {"images": ["", ...], "texts": ["", ...]}
9 """
```

Listing 3.1: Prompt for ChatGPT-4o and LLaMA-3:70b to analyze the HTML file of the Wikipedia article about Marie Curie. It provides additional information, stepwise task instructions and a template for the output.

The first attempt was conducted with ChatGPT-4o, that generated a response with the following content (see Listing 3.2). Importantly, it has to be respected that only the relevant lines are shown, because the original response was too large to be shown in the Listing.

```
{
    "images": [
        "https://upload.wikimedia.org/wikipedia/commons/thumb/3/31/
            Marie_Curie_c1920.jpg/220px-Marie_Curie_c1920.jpg",
        "https://upload.wikimedia.org/wikipedia/commons/thumb/7/71/
            Marie_Curie%2C_1903_Nobel_Prize_in_Physics.jpg/200px-
            Marie_Curie%2C_1903_Nobel_Prize_in_Physics.jpg",
        "https://upload.wikimedia.org/wikipedia/commons/thumb/0/05/
            Marie_and_Pierre_Curie.jpg/300px-Marie_and_Pierre_Curie.jpg
            ",
        "https://upload.wikimedia.org/wikipedia/commons/thumb/c/c9/
            Marie_Curie_lecture_in_Brussels.jpg/250px-
            Marie_Curie_lecture_in_Brussels.jpg",
        "https://upload.wikimedia.org/wikipedia/commons/thumb/2/23/
            Marie_Curie_and_her_daughters.jpg/250px-
            Marie_Curie_and_her_daughters.jpg"
    ],
    "texts": [
```

```
        "Marie łSkodowska Curie -(18671934) was a Polish and
            naturalized-French physicist and chemist who conducted
            pioneering research on radioactivity.",
        "As the first woman to win a Nobel Prize, Curie remains the
            only woman to win the Nobel Prize twice and the only person
             to win Nobel Prizes in two different scientific fields (
            Physics and Chemistry).",
        ...
        "During World War I, she developed mobile radiography units to
             provide X-ray services to field hospitals.",X
        ...
    ]
}
```

Listing 3.2: Generated response showing an analyzed HTML file as JSON via ChatGPT-4o. Taken from OpenAI, 2024b.

Thereon, another attempt was carried out via LLaMA-3:70b, using the same HTML file and composed prompt.

The following Listing (3.3) demonstrates the generated response by LLaMA-3:70b:

```
"""
{
    "images": [
        "https://upload.wikimedia.org/wikipedia/commons/c/c8/
            Marie_Curie_c._1920s.jpg"
    ],
    "texts": [
        "Marie Curie",
        "Polish and French physicist and chemist (1867-1934)",
        "...",
        "She was the first woman to win a Nobel Prize, the first
            person to win two Nobel Prizes in different fields (physics
             in 1903 and chemistry in 1911), and the first woman to
            become a professor at the University of Paris.",
        "...",
        "Early life and education",
        "Marie Curie was born in Warsaw, Poland on November 7, 1867.
            Her parents were both teachers, and she was the fifth of
            seven children.",
        "...",
        # All other texts about Marie Curie...
    ]
}
```

```
18 """
```

Listing 3.3: Generated response showing an analyzed HTML file of the Wikipedia article about Marie Curie as JSON via LLaMA-3:70b. Taken from Meta AI (2024a).

At first glance, the results seemed promising. However, a more precise evaluation led to the following results (see Table 3.2):

Description	ChatGPT-4o	LLaMA-3:70b
Response Time	5 to 10 seconds	9 to 10 minutes
Image URLs	Image URLs were parsed with hallucinations	Only one image was extracted correctly
Texts	Texts in line 10, 11 and 13 are incomplete (cut-off)	Apart from the heading, all parsed sentences were cut-off. A comment symbol (#) was added to the JSON that is forbidden by its syntax

Table 3.2.: Evaluation and comparison of the generated response by ChatGPT-4o and LLaMA-3:70b of the Wikipedia article about Marie Curie.

As discussed in Section 2.2.3, the web version of ChatGPT-4o has a fixed rate, so it does not charge for every written prompt and generated response. Due to this reason, ChatGPT-4o was used for this empirical and iterative research. However, for a better insight of the costs, which would have theoretically arisen, they were calculated by using the tokenizer[5] from OpenAI. The general price list is shown in Section 2.2.3 and the estimated costs for this approach with OpenAI API are presented below (see Table 3.3):

Description	Input Tokens	Output Tokens	Cost ($)
HTML file	186.750		0.466875
Prompt	136		0.00017
Response		541	0.00541
Total	186.886	541	∼0.47

Table 3.3.: Breakdown of used tokens and the estimated costs (USD) for analyzing the HTML file from Wikipedia article about Marie Curie with ChatGPT-4o.

The total costs for a single approach, as shown in Table 3.3, result in $0.47 by an amount of 187.427 tokens used. However, the realistic costs with the API would probably be much lower than $0.47 per prompt and response, because the

[5]https://platform.openai.com/tokenizer

context size (how many tokens fit into a single prompt) are limited to 128.000 tokens.

In the context of LLaMA-3:70b, it is important to take the costs for the required hardware into account (Shadow PC with 74.99€ per month). Apart from the hardware costs, the number of used tokens is irrelevant to be decoded here, because there are no rates for each input or output token used.

In conclusion it can be said, that the achieved results in both phases are not suitable for the requirements of the present project. The reasons are, that the extracted data were, firstly, parsed with either hallucinations or parts of the texts were cut-off, and secondly, the image URLs were partially made up. In particular, the shown performance in the experiment with LLaMA-3:70b was inappropriate, even though a Shadow PC was used.

In the next section, an investigation is undertaken using image recognition capabilities of the Large Language Models.

3.5.1.2. B: Analysis of Screenshot

In comparison to the text processing attempts in the previous sections, these attempts utilized the vision capabilities of ChatGPT-4o and LLaVA-3:34b for extracting contextually relevant texts and images from a screenshot. The screenshot, see Figure 3.4, was taken from the first part of the Wikipedia article about "Marie Curie".

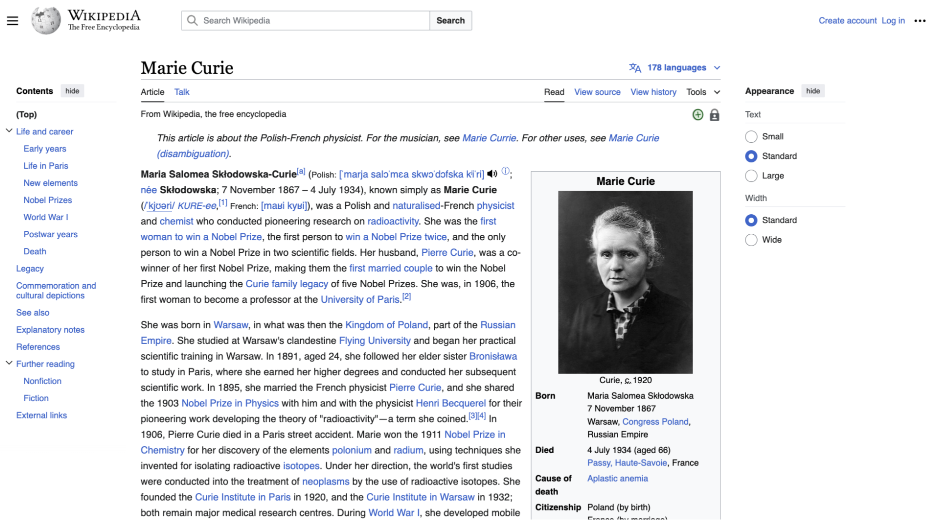

Figure 3.4.: Screenshot of the first fold of the English Wikipedia article about Marie Curie for ChatGPT-4o. Taken from Michaelsexton2003, 2024b.

The requirements for both attempts are listed below:

1. Texts must be of complete length (without cut-offs).
2. Texts mut not contain any spelling or wording mistakes (no hallucinations).
3. Images must directly be described, instead of picking their source URLs.
4. The filtered and extracted data must be collected in a JSON format, thus further processing in the development is easier.

Listing 3.4 shows the prompt, that is structured in the same manner as the prompts in the previous runs. The first line provides additional information about the context and line 3 to 5 contain a stepwise instruction how to treat the screenshot and how data should be extracted from it. Finally, line 7 to 9 present a JSON format that should be used to collect the data in a structured way.

```
1 Analyze the attached image. The image is a screenshot from the Wikipedia article about Marie Curie (english version).
2
3 Step 1 - Extract all texts from the screenshot, which refer to Marie Curie. Make sure to capture full sentences. Ignore: Logo, Banner, Searchbar, Navigation, Agenda, Buttons.
4 Step 2 - Describe all images regarding Marie Curie with 200 characters. Ignore: Logo, Icon, Navigation, Buttons.
5 Step 3 - Parse the information in the array in triple quotes. Don't answer anything else.
6
7 """
8 {"texts": ["...", ...], "images": ["...", ...]}
9 """
```

Listing 3.4: Prompt for ChatGPT-4o and LLaVA-3:34b to analyze the attached screenshot of the Wikipedia article about Marie Curie. It provides additional information, stepwise task instructions and a predefined pattern for the output.

Firstly, the generated response of ChatGPT-4o is shown in the Listing 3.5 below.

```
{
  "texts": [
    "This article is about the Polish-French physicist. For the musician, see Marie Currie. For other uses, see Marie Curie (disambiguation).",
    "Maria Salomea Sktodowska-Curie (Polish: '[marja salo'mea skwo' dofska ki'ri]), éne Sktodowska; 7 November 1867 - 4 July 1934), known simply as Marie Curie (/kjvari/ KURE€-, French: [maui kywi]), was a Polish and naturalised-French physicist and chemist who conducted pioneering research on radioactivity. She was the first woman to win a Nobel Prize, the first person
```

```
to win a Nobel Prize twice, and the only person to win a Nobel Prize in two scientific fields. Her husband, Pierre Curie, was a co-winner of her first Nobel Prize, making them the first married couple to win the Nobel Prize and launching the Curie family legacy of five Nobel Prizes. She was, in 1906, the first woman to become a professor at the University of Paris.",
    "She was born in Warsaw, in what was then the Kingdom of Poland, part of the Russian Empire. She studied at Warsaw's clandestine Flying University and began her practical scientific training in Warsaw. In 1891, aged 24, she followed her elder sister Bronistawa to study in Paris, where she earned her higher degrees and conducted her subsequent scientific work. In 1895, she married the French physicist Pierre Curie, and she shared the 1903 Nobel Prize in Physics with him and with the physicist Henri Becquerel for their pioneering work developing the theory of 'radioactivity'-a term she coined.",
    "In 1906, Pierre Curie died in a Paris street accident. Marie won the 1911 Nobel Prize in Chemistry for her discovery of the elements polonium and radium, using techniques she invented for isolating radioactive isotopes. Under her direction, the world's first studies were conducted into the treatment of neoplasms by the use of radioactive isotopes. She founded the Curie Institute in Paris in 1920, and the Curie Institute in Warsaw in 1932; both remain major medical research centres. During World War I, she developed mobile"
  ],
  "images": [
    "Marie Curie, c. 1920. Black and white portrait showing Marie Curie with short hair, looking directly at the camera, wearing a dark outfit."
  ]
}
```

Listing 3.5: Generated response showing an analyzed screenshot file as JSON via ChatGPT-4o. Taken from OpenAI, 2024c.

Secondly, the following Listing 3.6 illustrates the result generated by LLaVA-3:34b.

```
The image you've provided appears to be a screenshot of an email conversation or message. The content of the text
is not clear due to the resolution and angle at which the photo was taken, along with some text being obscured by
the phone itself in the way it was held when taking the picture. If there are specific details you're interested
```

```
in from the image, please let me know, and I can try to assist based on what is visible or provide guidance on how
to take a clearer screenshot.

>>>
>>> Analyze the base64 image. The image is a screenshot from the Wikipedia article about Marie Curie (english version).
The image appears to be a blurry or low-resolution screenshot of a computer screen displaying a webpage. It's not
clear enough for me to provide details about the content of the webpage, but it seems like it might be related to
historical figures given the context provided by the user.

>>>
>>> Step 1 - Extract all texts from the screenshot, which refer to Marie Curie. Make sure to capture full sentences. Ign
... ore: Logo, Banner, Searchbar, Navigation, Agenda, Buttons.
I'm sorry, but I cannot process this request as it is beyond my capabilities. As a text-based AI assistant, I do
not have the ability to analyze or extract information from images. To provide assistance with extracting text
from an image, you would need a specialized tool or service designed for optical character recognition (OCR).

>>> Step 2 - Describe all images regarding Marie Curie with 200 characters. Ignore: Logo, Icon, Navigation, Buttons.
I'm sorry, but I cannot process this request as it is beyond my capabilities. As a text-based AI assistant, I do
not have the ability to analyze or extract information from images. To provide assistance with describing an
image, you would need a specialized tool or service designed for visual recognition and analysis.<A picture of
Marie Curie holding a test tube. A portrait of her in scientific attire.

>>> Step 3 - Parse the information in the array in triple quotes. Don't answer anything else.
I'm sorry, but I cannot process this request as it is beyond my capabilities. As a text-based AI assistant, I do
not have the ability to parse or interpret data stored in an array format. To provide assistance with parsing
arrays, you would need a specialized tool or service designed for data manipulation and analysis.
```

Listing 3.6: Generated response showing an analyzed screenshot file as JSON via LLaVA-3:34b. Taken from Meta AI, 2024b.

The following Table 3.4 presents the evaluation and comparison of the image analysis processes. For better insights, the results of both attempts can be visually compared with the provided screenshot:

Description	**ChatGPT-4o**	**LLaVA-3:34b**
Response Time	5 to 10 seconds	Less than 9 minutes
Image Descriptions	The image was detected and described correctly	This process failed; the AI assumed that the image is an email, taken from a phone in a bad camera angle. In addition, it apologizes several times, that information could not be extracted.
Texts	Texts were extracted correctly, however the second paragraph was cut-off in the middle	Neither texts from the screenshot nor a JSON file were generated

Table 3.4.: Evaluation and comparison of the generated response by ChatGPT-4o and LLaVA-3:34b of the Wikipedia article about Marie Curie.

An additional highlight of ChatGPT-4o is that the AI successfully managed to only extract content from the main article and not anything else (menus, search bar, etc.)

The costs for images are calculated differently than for texts. For this theoretical estimation a vision pricing calculator[6] from OpenAI was used:

Description	**Input Tokens**	**Output Tokens**	**Cost ($)**
Screenshot (1950 x 1061)	1.105		0.005525
Prompt	123		0.00068
Response		529	0.0007935
Total	1.228	529	∼0.007

Table 3.5.: Breakdown of used tokens and the estimated costs (USD) for analyzing the screenshot of the Wikipedia article about Marie Curie with ChatGPT-4o.

As a result, the sum of used tokens by ChatGPT-4o (input and output) were 1.757 with total costs of $0.007. Assuming that the prompt remains identical throughout multiple usages, the costs can vary by different image resolutions.

[6]https://openai.com/api/pricing/

Even though ChatGPT-4o left out one paragraph in the second half, the overall result, involving reliability and costs, is very good. The only downside is, that there is no direct way to merge the generated data back into the DOM of the webpage, since no DOM references to the code are provided (e.g. IDs). Visual solutions such as prepending reference IDs in the rendered DOM would be very complex and especially unstable.

In conclusion it can be said, that this approach showed more potential than the previous one (see Section 3.5.1.2). All relevant elements were correctly detected, the image was considered as part of the article and almost all texts were correct. However, the missing possibilities to reference the corresponding elements in the DOM of the webpage made this approach unsuccessful.

3.5.1.3. Recap

	A: HTML file		**B**: Screenshot	
Description	**ChatGPT**	**LLaMA**	**ChatGPT**	**LLaVA**
Response Time	5 to 10 seconds	9 to 10 minutes	5 to 10 seconds	Less than 9 minutes
Correct Image URLs	Incorrect due to hallucinations	Single valid URL; others missing	-	-
Correct Image Descriptions	-	-	Image detected and described correctly	Failed
Correct Sentences	Texts were cut-off	Cut-offs and invalid JSON syntax	Sentences were cut-off	Failed
Costs	$0.47	equipment	$0.007	equipment

Table 3.6.: Comparison of HTML and screenshot processing with ChatGPT-4o as well as LLaMA-3:70b and LLaVA-3:34b.

The Table 3.6 provides an overview of the four documented attempts using Generative AI to extract images and text from web pages. In terms of performance, quality and costs, ChatGPT-4o performed best. Once again, it should be noted that LLaMA-3:70b and LLaVA-3:34b are free LLMs, but without specialized hardware, the computing time would be immeasurable.

Although ChatGPT-4o showed potential, the results were not reliable enough to use these extraction methods for the prototype development. In method A, many sentences were truncated or some words were creatively modified. Furthermore, some of the URLs of the images were freely invented. The screenshot method (B) led to better results overall. The texts were extracted correctly and the images were described appropriately, only the missing reference to the DOM of the Wikipedia article made further development with this method extremely complicated.

In contrast to that, an attempt was subsequently made to filter images and associated text from a webpage, based on JavaScript, which is described below.

3.5.2. Approach 2: Image Identification with Raycasting

Building on the previous Section's findings, the extraction of contextually relevant images and texts with Generative AI did not work reliably enough. In order to counteract on this, an alternative approach was tested.

The aim of this approach was to obtain images in combination with corresponding texts by only using JavaScript. Therefore, this extraction method was divided into two sequential steps: firstly, images had to be found, which are placed in the main content (e.g. in the article), and secondly, each of these images had to be paired with a text element. Such a text element had to have a meaning that mainly refers to the content of the picture. However, since this extraction method was a non-AI solution, the context awareness was not managed by LLMs, but instead it had to be "simulated" through raycasting. The desired result was to receive a collection of image-text pairs that can be used with the AI in further steps.

In the next Section the operating principle of raycasting and the approach to use it for gathering related images and texts will be debated.

3.5.2.1. Basic Concept

Raycasting is a method that is commonly used in game development. Coutinho (2023) explains raycasting as a computational method that is used in games or computer graphics to simulate light rays. In this method, a straight line (ray) is projected from a specific point in space in any specific direction. In case this line intersects other objects, various attributes can be determined. These involve the length of the line (distance to the object), its material properties, normal vector and so on.

Figure 3.5 serves for a better understanding of the explanation above. It shows an emitted ray from a point in space, which is a straight line, and it crosses other objects in the same space. The rectangular objects, on the top and in the middle of the Figure, have two intersection points each. As a result, if an intersection between the ray and an object occurs, this method can measure the distance between the emitting point and the object and additionally obtain various information from the object.

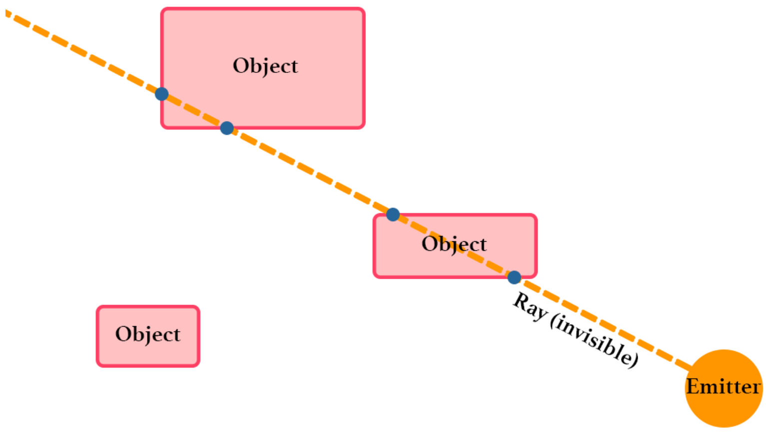

Figure 3.5.: Basic principle of raycasting where the orange circle, that represents the emitter, casts a ray (orange) over two objects (red rectangles). The four intersection points (blue) between the ray and the objects' boundaries are shown.

In this approach, the idea was to transform raycasting into a data collecting tool, that is able to retrieve image-text couples from rendered webpages[7]. It has to be highlighted, that the prerequisite is the assumption, that images and contextually related texts are physically located close to each other on the webpage.

In order to get insights in the application of raycasting on rendered webpages, the most important steps of the algorithm will be explained within the next sections.

3.5.2.2. Filtering of Images and Texts

To begin with, Figure 3.6 presents a section of a Wikipedia article, in its original state, that has some paragraphs and images (even though they are cut-off in the screenshot).

[7]The term *rendered webpage* refers to the visual representation of a website in the web browser. This is the version of a website that users can interact with.

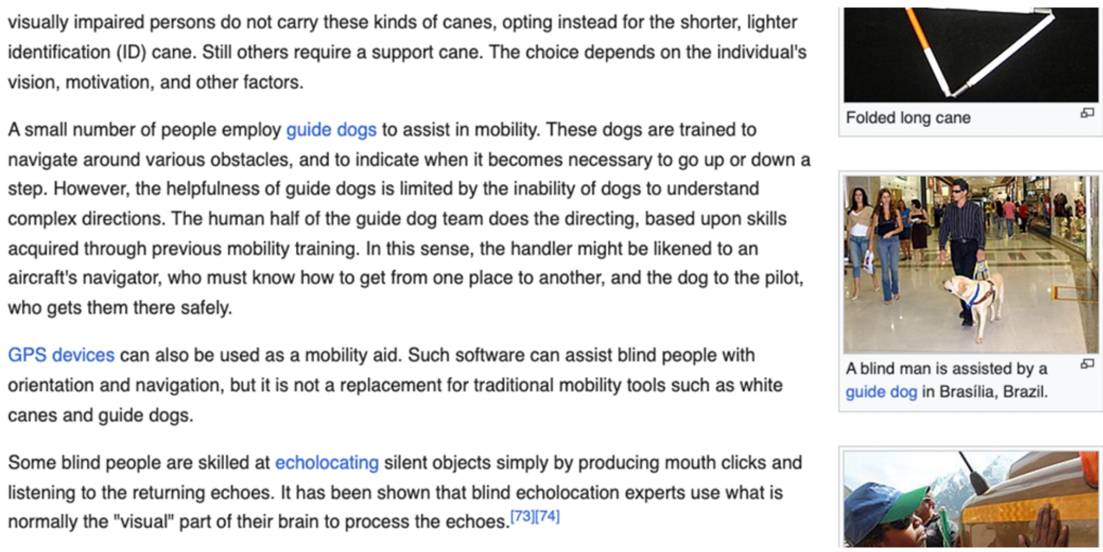

Figure 3.6.: Screenshot of the first fold of the Wikipedia article about Visual impairment. Taken from User-duck (2024).

In this step, the algorithm goes through all elements with the following HTML tags: paragraphs, headings and images. Obviously, the range of tags could be considerably extended, since images and texts can be implemented in many additional ways, however this would tremendously increase the complexity for this development.

As mentioned, this approach attempted to imitate context awareness, which would usually require LLMs. Therefore, the first part of context awareness is implemented by testing the above-listed HTML elements to fulfill certain criteria. These criteria are explained below:

- **Images:** It is assumed, that images with a greater rendered resolution than 80x80 pixels are rather used as images for content (e.g. as photos), than icons, logos or something else. In addition to that, in order to reduce upcoming problems, images with the extensions **.gif* (animated) or **.svg* (vector graphic) are excluded.
- **Headings:** These elements are taken by the algorithm straight away, because they are most probably only used for featuring content.
- **Paragraphs:** It is supposed that paragraphs appear in many different cases. In this approach, a text is defined as useful if it contains more than 8 words, so it builds a valid sentence.

As a result, Figure 3.7 presents the selection of texts and images according to the criteria above. As illustrated, the blue rectangles represent selected texts and the red rectangles show chosen images. However, the two captions underneath the upper and lower images are not picked, because their underlying HTML tag is a caption (not included in the search list).

In this Section, some images are shown with added highlights. These only serve as a visual approximation of the algorithm, but not as an exact technical construction (regarding pixel distances, colors, etc.).

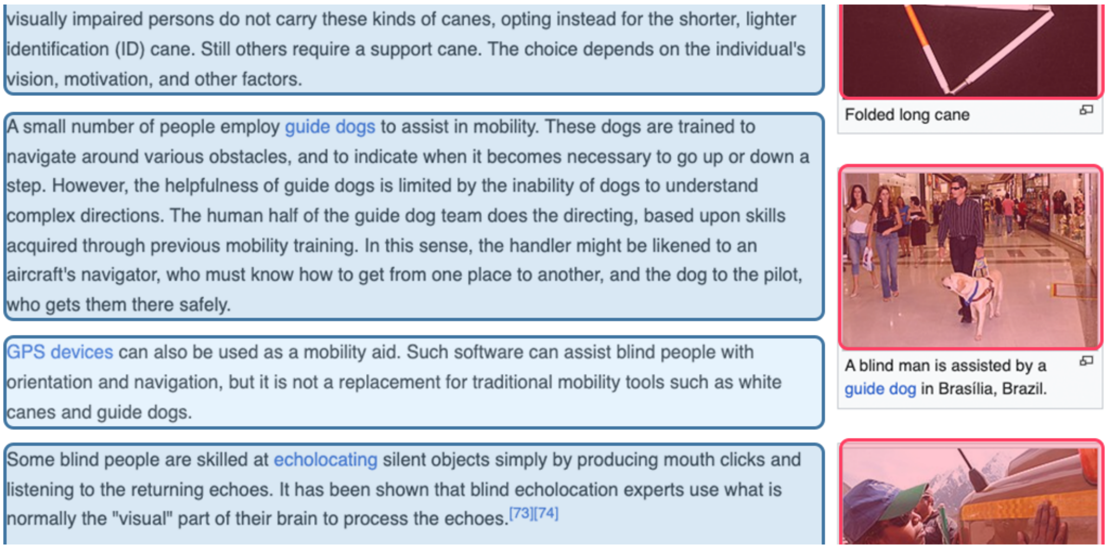

Figure 3.7.: Screenshot of the first fold of the Wikipedia article about Visual impairment. Selected images are highlighted in red, selected texts are highlighted in blue. Taken from User-duck (2024).

3.5.2.3. Calculation of Ray Intersections

Figure 3.8 illustrates the principle of raycasting on the basis of the red highlighted image. Point A, which is the coordinate of the image's center, has the role of the emitter. Starting from point A, a bunch of rays (red lines) are created. Red dots are located along the edges of the blue borders of the rectangles, that highlight the paragraphs. These red dots represent the intersection points between the ray and the boundaries of the paragraphs. According to this idea, intersections with other images are ignored, because an image can only be paired with a text element.

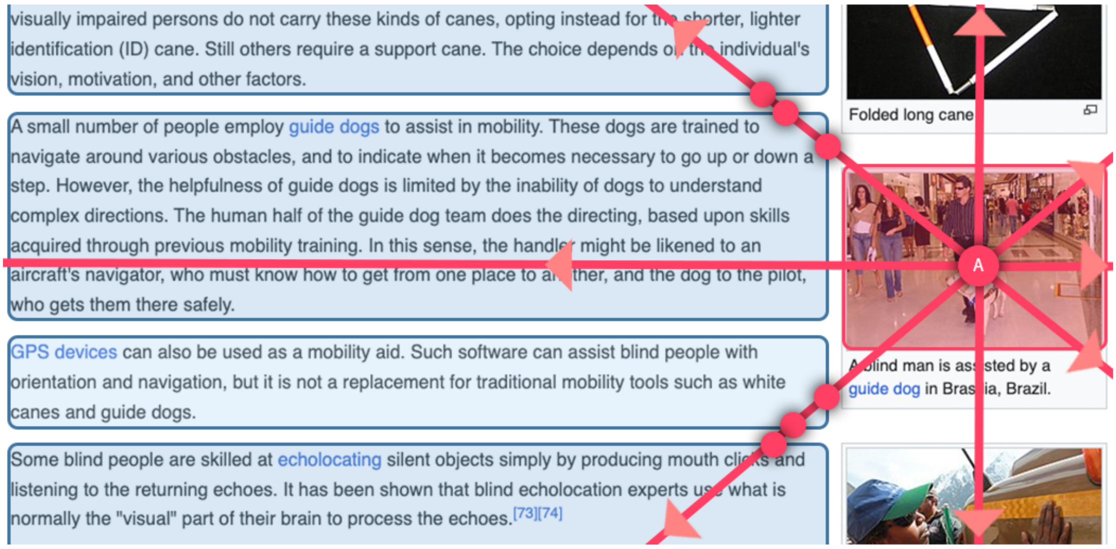

Figure 3.8.: The selected image (emitter) for this loop iteration is highlighted in red and rays are cast in all directions with 45° offsets. Taken from User-duck (2024).

Firstly, in order to enable paragraphs to be crossed with rays, they require boundaries. Figure 3.8 gives a cue how these boundaries were shaped in the prototype. However, the real ones are only a bit closer to the text and have no rounded corners.

Listing 3.7 presents the function `calculateEdges(...)` that is used to create four boundaries (edges) for any object provided as a parameter `element`. In the project, however, this function is only called on paragraphs. Thus, the method `getBoundingClientRect()` is utilized to obtain the paragraph's top-left corner point (*x/y*) and its size (width and height).

Given these values, as seen in Listing 3.7, lines 3 to 6 encompass the calculations of the boundaries' four corner points (*x/y*), that are subsequently mapped to edges (lines between two points) during lines 9 to 14. As a result, these edges appear in a rectangular shape.

```
1 function calculateEdges(element) {
2     const rect = element.getBoundingClientRect();
3     const corners = [
4         {x: rect.x, y: rect.y}, // top-left
5         {x: rect.x + rect.width, y: rect.y}, // top-right
6         {x: rect.x + rect.width, y: rect.y + rect.height}, // bottom-
             right
7         {x: rect.x, y: rect.y + rect.height} // bottom-left
8     ];
9     const edges = [
```

```
        { p0: corners[0], p1: corners[1] }, // Top edge
        { p0: corners[1], p1: corners[2] }, // Right edge
        { p0: corners[2], p1: corners[3] }, // Bottom edge
        { p0: corners[3], p1: corners[0] } // Left edge
    ];

    return {html: element, edges: edges};
}
```

Listing 3.7: JavaScript function to determine boundaries (edges) for any HTML element

Secondly, it is essential to emit a ray from a certain point (*x/y*) along a direction, that can possibly cross a boundary. Listing 3.8 shows the function-call of `calculateRayEndpoint(...)` that casts a line from any `startPoint` (*x/y*) in an arbitrary angle `degree` with an exact distance of `500` pixels. Due to these given parameters an exact end point (*x/y*) for the ray is returned. It should be noted, that the increment of `10` (line 1) is used to save computing power.

```
1 for (let degree = 0; degree < 360; degree += 10) {
2     const endPoint = calculateRayEndpoint(startPoint, degree, 500);
```

Listing 3.8: Loop that calculates the 36 ray end points in a circle.

Finally, the core of this implementation is to detect intersection points, that may appear between rays and boundaries. Therefore, Listing 3.9 shows the function `lineIntersection(...)` that is used with the parameters outlined below:

1. **p0:** ray start (emitter) point (*x/y*).
2. **p1:** ray end point (*x/y*).
3. **p2:** edge start point (*x/y*).
4. **p3:** edge end point (*x/y*).

This function allows to identify a possible intersection point (*x/y*) between two lines, namely a ray and an edge, at a time. Hence, it has to be used with the same ray for four times per paragraph for seeking an intersection, because a paragraph has four boundaries. In case an intersection point (*x/y*) was found, it will be returned.

```
function lineIntersection(p0, p1, p2, p3) {
    const s1_x = p1.x - p0.x;
    const s1_y = p1.y - p0.y;
    const s2_x = p3.x - p2.x;
    const s2_y = p3.y - p2.y;

    const s = (-s1_y * (p0.x - p2.x) + s1_x * (p0.y - p2.y)) / (-s2_x
        * s1_y + s1_x * s2_y);
    const t = ( s2_x * (p0.y - p2.y) - s2_y * (p0.x - p2.x)) / (-s2_x
        * s1_y + s1_x * s2_y);
```

```

    if (s >= 0 && s <= 1 && t >= 0 && t <= 1) {
        return {
            x: p0.x + (t * s1_x),
            y: p0.y + (t * s1_y)
        };
    }

    return null;
}
```

Listing 3.9: JavaScript function that takes two lines and returns an intersection point if an intersection occurs

In the end, for all filtered images that are looped in the implementation, each iteration brings an individual circle of 36 rays onto the webpage with corresponding intersection points.

3.5.2.4. Detection of Shortest Distance

This is the final step to imitate the context awareness on a webpage. Therefore, all images with their series of intersection points are looped. Within the scope of each image, the underlying intersection points are compared with each other regarding their distance to the ray-emitting center point A of the image. As a result, the closest point B is returned and considers the paragraph as contextually relevant to the image.

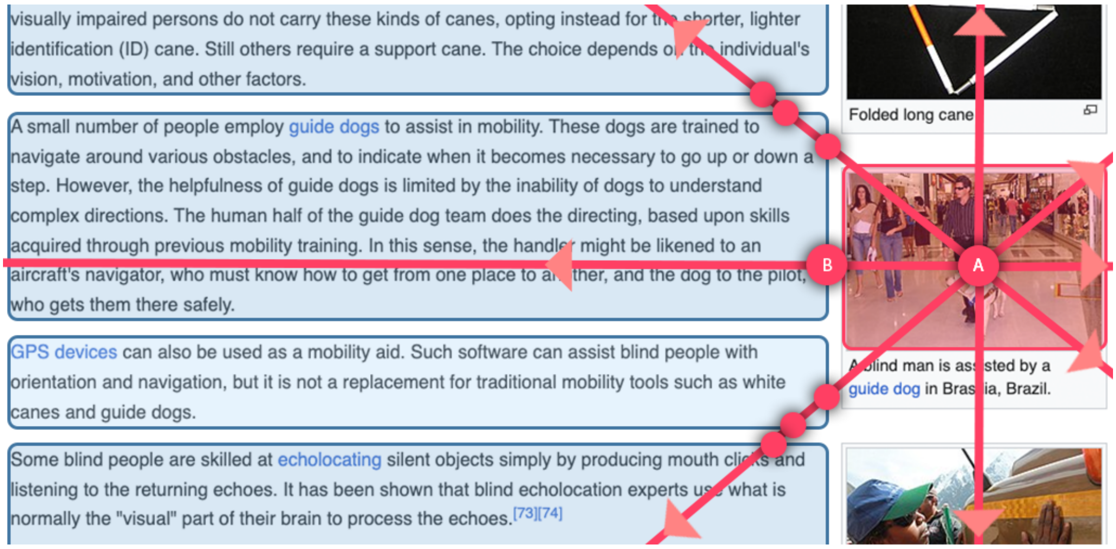

Figure 3.9.: Selected images are highlighted in red, selected texts are highlighted in blue. The shortest distance (A and B) are shown.Taken from User-duck (2024).

Once the relevant image-text pair has been identified by this algorithm, the data is passed on to the AI. Various attempts to create suitable image descriptions from the data obtained are discussed in the next section.

3.5.3. Approach 2: Image Descriptions with AI (Part 2)

At this time in the research it became clear, that LLaMA-3:70b and LLaVA-3:34b cannot continue to be used for this development. The reason is, that Shadow strictly forbids using their virtual machines as servers (Shadow, n.d.). However, such a server would be necessary for the prototype to run with a certain performance, that leads in a proper response quality. Furthermore, utilizing a dedicated server would be technically feasible, but due to the private funding it would exceed the budget.

Given the reasons above, the following experiments were conducted using ChatGPT-4o and the Assistants API on the OpenAI platform. ChatGPT-4o was utilized in order to save costs during the examination. In order to achieve a comparable overview, the three attempts, divided into three sections, were schematically carried out using the same image resource and corresponding input text each —both taken from the Wikipedia article about "Visual impairment" (as used in the raycasting section). However, it has to be taken into account, that the debated approaches during the following sections are the most relevant ones for the comparison of describing images with AI. While some approaches had significant differences, others only involved some little adjustments.

In preparation for this experiments, a list containing category names and related questions was created with the aid of ChatGPT-4o. This category list allowed the AI to focus on certain image features while producing image descriptions. The outcome is presented below:

- **Scenery:** How does the environment look like? Is it a city or landscape? Is there water, snow, ice, grass or sand?
- **People:** Who are they? What is the interaction or situation? What are their body and facial expressions?
- **Cultural:** Why is the situation remarkable? What is the background information?
- **Animals:** Pet, wildlife, marine life or birds? What color are they? What are their interactions?
- **Objects:** Household, items, technology, vehicles, good and drink, clothing and accessories, art and craft? How big is it? Does it have a certain name or brand name? Is it old or new?
- **Travel:** Which type of travel? Cars, trains, airplanes, boats? What about the tourism? Is it a certain tourist spot?
- **Architecture:** Is it a residential like a house or apartment? Is it a commercial building? Or is it historical like a castle, ruin, monument or religious building like a church, temple or mosques?
- **Emotional and Abstract:** What is the mood (e.g. happy, sad, angry)? Is it like modern art, surreal scenes, or else? Are any signs included?
- **Time and Seasons:** Day and night, What seasons? Is it Christmas, Halloween, New year or something like that?
- **Historical and Cultural:** What event was it? A war, revolution, or else? It is a traditional festival, ritual, costumes or a historical figure?

It is obvious, that the provided range of chosen categories can only cover a tiny range of contexts of image features. However, this information was still suitable enough to work at a proof-of-concept. See the following Sections for a more detailed explanation of how this list was used.

3.5.3.1. Multiple Prompts with Categories

This Section debates an attempt using ChatGPT-4o in a conversational style, which means that multiple prompts were submitted to the AI. This run consisted of two cycles, where the first iteration is explained as follows:

Message 1:
The goal of the first prompt was to make the AI figure out which categories can be matched with the given image (see Figure 3.10).

Figure 3.10.: Picture of the blind man with a guide dog. Taken from Cruz (2006).

After attaching this image to the ChatGPT-4o chat, the prompt shown in Listing 3.10 was prepared. This prompt contains a text passage from the original Wikipedia article, presented in quotes in line 1. Next, the category names, without the according questions, are provided in triple quotes from lines 3 to 5. Lastly, lines 7 and 9 show instructions for the AI to try to determine which category names (lines 3 to 5) fit best to the content of the attached image. It was assumed, that ChatGPT-4o also takes the attached text in line 1 as a reference too.

```
1 Find the image attached to the context "A small number of people
    employ guide dogs to assist in mobility. These dogs are trained to
     navigate around various obstacles, and to indicate when it
    becomes necessary to go up or down a step.".
2
3 """ ["scenery", "people", "animals", "objects", "travel", "
4 architecture", "art", "emotional", "time and seasons", "historical
5 and cultural", "health", "business"] """
6
7 Step 1 - Analyse the image and check if it matches with one or more
8 categories in the array in triple quotes.
9 Step 2 - Return the selected categories as array and 'DONT write any
    other text.
```

Listing 3.10: Prompt for ChatGPT-4o to analyze the picture of the blind man with a guide dog, provided additional information and predefined categories as array.

As expected, the generated reply contained an array with three chosen categories (see Listing 3.11). In fact, it was not guaranteed that the outcome would be formatted as desired, it could have also been expressed as natural text or something else.

```
["people", "animals", "health"]
```

Listing 3.11: Generated response showing an array with selected categories by ChatGPT-4o. Taken from OpenAI (2024a).

Message 2:
As a second step, ChatGPT-4o was requested to create an image description based on the specific questions, that were created together with the original categories list. Their purpose was to put the focus of the image analysis into a certain contextual direction. The prompt for doing so is exhibited in Listing 3.12.

```
People: Who are they? What is the interaction or situation? What are
    their body and facial expressions?

Animals: Pet, wildlife, marine life or birds? What color are they?
    What are their interactions?

Health: Is it a hospital? Are their doctors? What kind of doctors? Is
     there any medical equipment?

Step 1 - Write an image description with the length of 300 characters
max. Be precise, brief and create only one paragraph.
Step 2 - Answer in english and format the answer as array ["Answer
text goes here ..."]. 'DONT write any other text.
```

Listing 3.12: Prompt for ChatGPT-4o to create a detailed image description for the picture with the blind man with a guide dog based on the questions.

After waiting a few seconds, the AI responded with an array containing the image description (63 tokens). This array is demonstrated in the next Listing (see 3.13):

```
["A visually impaired man is walking in a mall guided by a yellow
    Labrador wearing a harness. The man wears sunglasses, holding the
    'dogs handle confidently. People around are casually walking, some
     glancing at them. The dog, a trained guide, helps the man
    navigate the crowded space safely"]
```

Listing 3.13: Generated response based on the questions about the blind man with a guide dog by ChatGPT-4o. Taken from OpenAI (2024a).

Conclusively, the provided Table 3.7 presents a comparison of the costs of this attempt. Therefore, the used tokens for the image processing as well as both, prompts and responses (message 1 and message 2), are listed.

Description	Input tokens	Output tokens	Cost ($)
Image (1851 x 1287)	1105		0.002763
Prompt 1 & Response 1	136	9	0.00043
Prompt 2 & Response 2	120	59	0.00089
Total			0.004083

Table 3.7.: Breakdown of used tokens and the estimated costs (USD) for a multi-message approach.

In the following section, a similar attempt is discussed using a single prompt style.

3.5.3.2. Single Prompt with Categories

As mentioned in the previous attempt, there is no promise yet when using AI, that the generated response code is formatted as desired (correct syntax etc.). Such an uncertainty can create a risk for software with interfaces, when they receive data that do not match their format, that they crash.

The goal of the present approach was, firstly, to prevent generated coding errors, and secondly, to make the prompt shorter in order to save tokens. In order to make the comparison easier, the exact same image (see Figure 3.10) and context information were used again in this turn.

The picture was attached to the prompt, that is shown in Listing 3.14. This time, the category list was added to the prompt from lines 3 to 11. It has to be mentioned, that the code from line 9 on is removed from the Listing, because the complete list is too large for this demonstration.

```
1 Find the image attached to the context "A small number of people employ guide dogs to assist in mobility. These dogs are trained to navigate around various obstacles, and to indicate when it becomes necessary to go up or down a step.".
2
3 """
4 categories = [
5 "Scenery": "How does the environment look like? Is it a city or landscape? Is there water, snow, ice, grass or sand?",
6 "People": "Who are they? What is the interaction or situation? What are their body and facial expressions?",
```

```
"Cultural": "Why is the situation remarkable? What is the background
    information?",
"Animals": "Pet, wildlife, marine life or birds? What color are they?
     What are their interactions?",
...
]
"""

Step 1 - Analyse the image and check if it matches one or more
categories (keys) in the array in triple quotes.
Step 2 -Describe the image based on the values of selected keys.
Step 3 -Write a description in the length of max 400 words. Parse
this description as string in the array ‘“’...[] and ’DONT reply
with any additional text.
```

Listing 3.14: Prompt for ChatGPT-4o to analyze the picture of the blind man with a guide dog based on the provided information and categories and thus creating an image description.

After a few seconds, ChatGPT-4o returned a image description in natural language, nested in an array, shown in Listing 3.15. It is evident, that this generated array is much longer than the previous one (180 tokens). The reason is, that the maximum length, requested in the previous prompt, was set to 400 words. The value of the maximum length was adjusted, in order to figure out whether a longer text automatically describes more relevant image features.

```
["The image depicts a visually impaired man walking inside a shopping
     mall with the assistance of a guide dog. The man is wearing dark
    sunglasses and is holding onto the harness handle attached to his
    guide dog, a yellow Labrador Retriever, which helps him navigate
    the crowded space. The mall is busy with several people casually
    walking around, some carrying shopping bags. The guide dog is
    leading the way confidently, indicating that it is well-trained
    for mobility assistance. The scene highlights the use of guide
    dogs for individuals with visual impairments to safely move around
     public spaces. This image can be categorized under ’People’ due
    to the focus on the interaction between the man and his guide dog,
     ’Animals’ since it features a service animal in action, and ’
    Cultural’ as it demonstrates an important aspect of accessibility
    and independence for the visually impaired in society."]
```

Listing 3.15: Generated response about the analyzed picture of the blind man with a guide dog containing the image description as array. Taken from OpenAI (2024a).

Finally, Table 3.8 presents the estimated used tokens and occurred costs of this attempt.

Description	Input tokens	Output tokens	Cost ($)
Image (1851 x 1287)	1105		0.002763
Prompt	493	144	0.0026725
Total			0.0054355

Table 3.8.: Breakdown of used tokens and the estimated costs (USD) for creating the image description about the blind man with a guide dog with ChatGPT-4o.

In particular, the final attempt uses a special tool from OpenAI in order to eventually reduce the amount of used tokens even more. Details will be documented in the next section.

3.5.3.3. Assistants API with Categories

So far, in every ChatGPT-4o request for describing an image, the prompt was adjusted and the entire categories list was included. In fact, this way leads to costly redundancies, because the category list, which stays the same, is submitted to the AI multiple times and consuming tokens.

Consequently, the aim of this approach was to stick to a robust single-prompt conversation style, however without sending the category list that often. As a solution, OpenAI provides a tool called the Assistants API. This is a service that enables developers to build their own chat bots for their custom applications. These chat bots provide options to be fine-tuned, to adopt a personality and apply specific skills (*Assistants API overview*, n.d.). During this attempt, the capabilities were used to upload the categories list as **.json* to the platform directly. As a result, the categories list can be excluded from prompts, because it can be accessed by writing proper instructions. As used in the former approaches, this attempt also uses the image in Figure 3.10 as a resource.

Moreover, the following Listing 3.16 presents the prompt that was used for this examination.

```
Find the image attached to the context "A small number of people employ guide dogs to assist in mobility. These dogs are trained to navigate around various obstacles, and to indicate when it becomes necessary to go up or down a step.".

Step 1 - Analyse the image and check if it matches with one or more categories (see keys in JSON file).
Step 2 - Take the questions (values of these keys) and write an image description based on them (and the context).
Step 3 - Description properties:
```

```

- max 300 words
- answer in one single paragraph
- use language "en"
- the answer text must not contain code, enumerations, bullet points,
    keys of the json, questions
- do not reply with any additional text.

Take your time!
```

Listing 3.16: Prompt for the Assistants API to create an image description, using the category list by referencing the categories.json file.

After the request was submitted to the Assistants API, a very detailed image description got returned (301 tokens). This image description is demonstrated in Listing 3.17.

```
In a bustling shopping mall, a visually impaired man navigates
   confidently with the help of his guide dog, a large, light-colored
    Labrador Retriever. The man, wearing dark sunglasses and dressed
   in a black striped shirt and dark trousers, holds onto the dog's
   harness, indicating his reliance on the animal for mobility. The
   guide dog, equipped with a harness that signifies its role as a
   service animal, leads the man through the busy environment,
   showcasing the trust and bond between them. The mall is lively,
   filled with various shops and casually dressed shoppers, some
   carrying bags, reflecting a typical day of shopping. The guide dog
    expertly maneuvers around obstacles, ensuring the man's safety
   and embodying the essential role these animals play in providing
   independence to those with visual impairments. This scene captures
    a moment of autonomy facilitated by the guide dog, emphasizing
   its importance in enhancing the quality of life for its handler.
   The image illustrates the practical application of guide dogs in
   real-world settings, demonstrating their ability to navigate
   complex environments and assist their handlers in moving safely
   and confidently. The presence of the guide dog not only aids in
   mobility but also symbolizes the empowerment and independence that
    service animals provide. This interaction highlights the critical
    role these animals play in everyday life, allowing their handlers
    to engage more fully and independently with the world, despite
   the challenges posed by visual impairments.
```

Listing 3.17: Generated response showing the resulting image description about the picture created by the Assistants API with a categories.json file. Taken from OpenAI (2024a).

A breakdown of the costs for this attempt is given in Table 3.9. As visible in this analysis, the prompt consumed less tokens with GPT-4o than the generated response, resulting in approximately $0.006 per AI request. Obviously, the costs can slightly vary due to image sizes and response lengths.

Description	Input tokens	Output tokens	Cost ($)
Image (1851 x 1287)	1105		0.002763
Prompt	158	267	0.003065
Total			0.005828

Table 3.9.: Breakdown of used tokens and the estimated costs (US-Dollar) with the Assistants API.

The next section provides a general overview of these three attempts.

3.5.3.4. Comparison of GPT-4o Approaches

All reported approaches were aimed to generate the best possible image descriptions. By doing so, it should not only describe the images in detail, but also involve relevant context information from the adjacent webpage. In addition to generating proper image descriptions, other factors such as the robustness as a software interface as well as keeping the costs low had to be considered. Therefore, the following Table (see Section3.10) serves as a comparison of the three approaches. However, to make the three attempts more comparable, the output tokens in the Table were set to 100 as the control value. This makes it possible to determine which of the three attempts is the most cost-effective in terms of its input variant. The actual output tokens used in the approaches can be found in the cost tables of the respective Sections.

Description	Context Quality & Output Length	Robust as API	Used Tokens	Costs ($)
Multi-Prompt	All described details of the image are completely related to the provided context details. (100 tokens)	-	1.461	0.0044
Single-Prompt	The analysis of the picture is suitable and matches the context. However, there are some extra sentences, that could have been ignored. (100 tokens)	✓	1.698	0.0049
Assistants API	Details about the context are accurately covered, but other irrelevant details too. (100 tokens)	✓	1.363	0.0041

Table 3.10.: Comparison of approaches based on description, context quality & output length, API robustness, total tokens used, and costs.

The Table 3.10 shows that the multi-prompt method achieved the best image description. The reason for this is probably that split prompts can be processed better by the AI. However, this approach is not robust enough with regard to the software interface. In contrast, the Assistants API proved to be more robust and cost-effective, although the relevant details for the image descriptions were not selected with the same level of precision.

Ultimately, it can be said that the Assistants API approach is the most suitable for this project. This approach makes it possible to work with only one prompt and one response per image, which enhances the software robustness, the category list can be updated at any time centrally on the OpenAI platform, which prevents redundant inputs, and the costs are proportionally the lowest. The only compromise that has to be made is that the image descriptions are generated with less contextual awareness.

3.5.3.5. Recap

Overall, approach 2 part 2 (3.5.3) delivered satisfactory results. The raycasting method can be considered rather unconventional for websites, but despite of some compromises it worked better than expected. These compromises arose from the fact that LLMs were not integrated for this method.

The research part with ChatGPT-4o produced a bunch of results that often only diverged in small details. It was always important to include as much information as possible, while saving as many tokens as feasible to keep costs as low as possible. Further experimentation with other parameters, such as the length of texts or the naming of categories, would be necessary. In any case, the category list for these approaches covered only a small part of possible image details and contexts, which was only designed for a proof of concept.

3.6. Implementation

In the previous Section, appropriate methods for extracting images and text from webpages as well as generating contextual image description were determined. In summary, this research resulted in choosing the raycasting method, which originally comes from game development, and the Assistants API, as this is the best suitable method.

This Section serves as a documentation of the realization of the prototype. The focus is on the specific features that make this prototype unique. All other development steps, such as the implementation of basic communication interfaces, are not explained. It should be noted that certain Sections of the previous Section were implemented in the course of the research.

This part includes a detailed description of the final software architecture, the development of a Chrome Extension, the development and deployment of an Express.js application as well as the design of the user interface.

3.6.1. Software Architecture

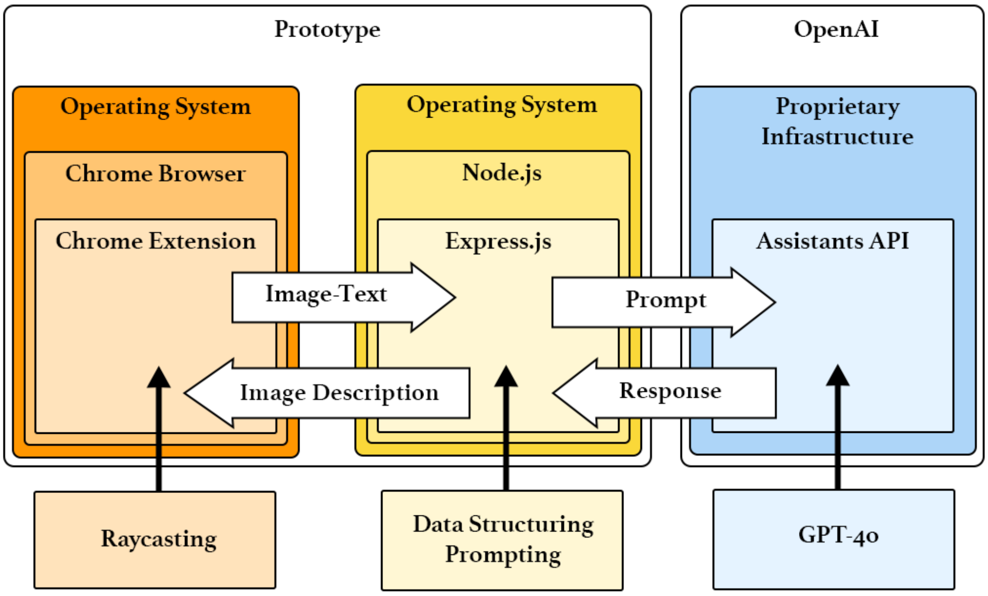

Figure 3.11.: Software architecture of the implemented prototype for describing images on webpages based on their context. The figure shows the three modules: the Chrome Extension, the Express.js app (middleware) and the third-party Assistants API.

Figure 3.11 illustrates the architecture design of the developed prototype. As noticeable, the prototype itself consists of two components, the Chrome Extension (orange area) and an Express.js app (yellow area). The Chrome Extension is an add-on for the Chrome Browser, so it works on any operating system that has the Chrome Browser installed. In addition to that, the Express.js app can be deployed to any local machine or server, the only prerequisite is that Node.js is set up. The LLM GPT-4o is integrated into the architecture by using the Assistants API (see blue area). This API is a third party service and can be accessed at any time.

The data flow proceeds as follows: After the user has initiated the procedure by pressing a shortcut, first, the Chrome extension uses the raycasting method to collect the relevant images and texts from a webpage and sends it to the Express.js app. Express.js structures the data and wraps them in a prompt along with further instructions, and then forwards them to the Assistants API. The Assistant's task is to create the image descriptions using GPT-4o and returning them to the Express.js app. Once the image descriptions have been received, they are finally passed back to the Chrome Extension to be integrated into the relevant places on the webpage.

The used components, Chrome Extension and Express.js, are described in more detail in the next Sections. Additional relevant implementation details that were necessary for seamless operation are also discussed.

3.6.2. Google Chrome Extension

Chrome Extensions have the purpose to access web pages and modify their DOM data. Since they are based on web technologies such as HTML, CSS and JavaScript, standard web development knowledge is a proper prerequisite. The goal of this Section is to present the most important development steps and decisions, the detailed documentation for building a Chrome Extension can be found in the Google documentation (Google, 2022).

As it can be seen in the following file structure (see Figure 3.12), the Chrome extension was named *Altify*. This product name is a combination of *alternative text* and *modify*.

```
altify
├── scripts
│   └── content.js
├── styles
│   └── fix.css
├── background.js
├── icon48.png
└── manifest.json
```

Figure 3.12.: File structure of Chrome Extension.

The two crucial files for this project are `content.js` and `manifest.json`. An essential component is the manifest, because it contains various information such as name, version, permissions and modules, but also the definition of shortcuts that can execute certain functions in the code when pressed (see Listing 3.18). The key `mac` stands for a mac specific shortcut and `default` is used on Windows.

```
"commands": {
    "image-to-text": {
        "suggested_key": {
            "mac": "MacCtrl+1",
            "default": "Ctrl+Shift+1"
        },
        "description": "Process image and text pairs with AI."
```

```
    }
}
```

Listing 3.18: Commands key in manifest.json for calling functions in code.

As soon as the mentioned key combinations have been pressed, in this case either on Macintosh or Windows, the raycasting algorithm (described in Section 3.5.2) is executed in `content.js`. An important detail is that each HTML element that is selected by the algorithm is assigned with a custom ID. These are `imageId` for the selected images URLs and `textId` for the associated text elements. These IDs are retained for the rest of the data processing in order to insert the generated image descriptions back into the correct place on the webpage at the end of the procedure.

Image URLs are grouped together with texts and their custom IDs in JavaScript objects and sent individually to Express.js. This object is exhibited in Listing 3.19. It has to be mentioned, that the functionality for communicating with Express.js is implemented in `background.js`, that runs in another thread.

```
{
    imageId: image.dataset.imageId,
    textId: text.dataset.textId,
    imageSrc: image.currentSrc,
    text: text.innerText
}
```

Listing 3.19: Data structure for image URLs and text pairs with custom IDs.

The next section explains how the data is further processed in Express.js so that it can ultimately be sent to the AI.

3.6.3. Backend

This framework is based on Node.js and is considered as a lightweight and flexible middleware solution. This means, that it provides stable capabilities for general web development, especially routing features for network traffic. However, the current Section only shows the important parts of the implementation used for the prototype, thus the necessary development steps to run this application is shown in the documentation (OpenJS Foundation, n.d.-a). Due to the provided usability, this framework was chosen in order to manage the traffic between the Chrome Extension and the Assistants API.

Node.js is a software that is used to run many different JavaScript applications on it, such as Express.js. This can be installed on the local computer or on a server, and it is free, open-source and cross-platform (it runs on Macintosh, Windows and Linux (OpenJS Foundation, n.d.-b).

The most important element of the Express.js application is the prompt, that sends the an image URL and text to the AI to generate an image description. This is shown in the Listing 3.20, whereby the exact explanation of the instructions is provided in Section 3.5.3.3. Essentially, two JavaScript objects are delivered from the Chrome Extension that are `imageAndText` and `metadata`. The first one, which is `imageAndText`, contains the imageURL, the text and both assigned IDs. Furthermore, the other object `metadata` provides additional information about the webpage. This involves the language code of the webpage and attributes from the meta tags. Both are used together in every prompt to provide as many context details as possible. The variable `imageAndText.text` is exhibited in line 2, that holds the corresponding text from the image (it is truncated to 500 characters to save tokens). Thereon, the variable `metadata.lang` is applied in line 8, that holds the used language from the webpage, in order to write the image description in the same language.

```
1  const apiPrompt = ‘‘ +
2      ‘Find the image attached to the context "${imageAndText.text.
          substring(0, 500)}".‘ +
3      ‘Step 1 - Analyse the image and check if it matches with one or
          more categories (see keys in JSON file).‘ +
4      ‘Step 2 - Take the questions (values of these keys) and write an
          image description based on them (and the context).‘ +
5      ‘Step 3 - Description properties: ‘ +
6      ‘- max 400 words‘+
7      ‘- answer in one single paragraph‘+
8      ‘- use language "${metadata.lang}"‘ +
9      ‘- the answer text must not contain code, enumerations, bullet
          points, keys of the json, questions‘ +
10     ‘- do not reply with any additional text.‘ +
11     ‘Take your time!‘;
```

Listing 3.20: Prompt that instructs AI to generate an image description from a given image URL and context text.

Since this prompt is written and submitted via JavaScript and not created via the ChatGPT web interface, images must be attached in a different way. This is presented in the next Listing (see 3.21) in lines 9 and 10. The object `imageAndText` is the same as mentioned before, that holds data such as the image URL. Express.js required the *OpenAI Node API Library*[8] within the node modules, in order to enable the communication. Therefore, it was also important to generate a unique *API-Key*, however this cannot be demonstrated in this paper, due to security reasons.

[8]https://www.npmjs.com/package/openai/v/4.0.0-beta.4

```
{
    role: "user",
    content: [
        {
            "type": "text",
            "text": `${apiPrompt}`
        },
        {
            "type": "image_url",
            "image_url": {"url": `${imageAndText.imageSrc}`}
        }
    ]
}
```

Listing 3.21: JavaScript code to attach an imageURL to a prompt.

Another very important aspect of the prototype development was to build a suitable infrastructure to make the project portable. It was mentioned earlier that the Express.js application can be run locally or on the server. For this purpose, a virtual private server (VPS) was rented from the provider world4you[9]. The specifications of the server infrastructure are outline below:

- **VPS:** Linux vServer S for 9.50 € per month. Further specifications can be found on the vendor website[10].
- **Operating System:** Debian.
- **Domain:** "contextbasedimages.cloud" with free SSL encryption.

3.6.4. User Interface

While the image description process is running, which takes a couple of seconds, it is always important to keep blind users informed of the progress. This is an essential element for guaranteeing the user satisfaction and also includes notifying the user if the description of an image has failed.

In the case of this prototype, it was particularly important that this progress information is communicated in an accessible manner that can be perceived by screen readers or braille displays. For ensuring this functionality, attributes known as ARIA attributes were used. ARIA stands for Accessible Rich Internet Applications and has the purpose, to provide HTML elements with additional information, that support the screen reader. In other words, they provide context details to the DOM, so blind users can better identify the element they focused

[9] https://www.world4you.com/
[10] https://my.world4you.com/en/shop/server

on (Firth, 2019). There are many different ARIA attributes, but the three ones used in the prototype are listed below (Firth, 2019):

- **aria-role:** Provides information representing the focused section, such as banner, main, navigation and so on.
- **aria-hidden:** Hides an element, that is only used for e.g. visual decoration.
- **aria-live:** Creates real-time notifications when sudden changes of the content appear.

The attributes just explained were used to display a status message as soon as the progress of the image descriptions has changed. In this case, `aria-role` was set to `alert`, `aria-hidden` to `false` and `aria-live` to `assertive`. In addition, a function has been implemented that defines the status message. This is shown in the next Listing (see 3.22). The variable `count` holds the actual number of images that were successfully described and `limit` shows the maximum amount of possible descriptions for a webpage. The limitation was implemented for testing purposes to reduce the number of described images, in order to save tokens.

```
heading.textContent = ‘Processing Status: ${count} of ${limit}‘
```

Listing 3.22: JavaScript code to communicate the image description progress status to assistive technologies.

3.6.5. Recap

In summary, it can be said that developing a prototype intended for blind people requires a more specialized way of thinking in some aspects. An example is, that user interfaces for blind people have different important factors, than typical graphical user interfaces, used by sighted people. While visual elements can be used as intended or even as decoration, interfaces for blind people have to be accurately integrated in the correct HTML elements.

Furthermore, the basic infrastructure was easy to create because various tools already offer many options, such as manipulating web pages with Chrome Extensions or sending prompts to the AI via JavaScript. These mentioned features, and many more, allowed the development of an automated image description algorithm.

In the next Section, the resulting prototype is demonstrated in action.

3.7. Demonstration

The prototype, in its current form, is presented in this section, accompanied by a discussion on meeting the specified requirements. The defined functional and non-functional requirements in Section 3.4 were used as a guideline to accomplish the objective of this thesis.

Wikipedia was an ideal medium for a proof-of-concept, due to the DOM tree and the many paragraphs and associated images, so the prototype was optimized for it. In addition to that, it was not important for the prototype to be a generic tool, that operates properly for many different websites, instead it serves as a proof-of-concept blind people should profit from.

It starts with a Wikipedia article in its original form, that is shown in Figure 3.13). As already used in former Sections, the English page about "Marie Curie" suited this presentation properly, however it got selected by coincidence. The important factor was that the webpage has an image and paragraph within the first fold. It should be noted, that the demonstration was limited to the first three images to save costs.

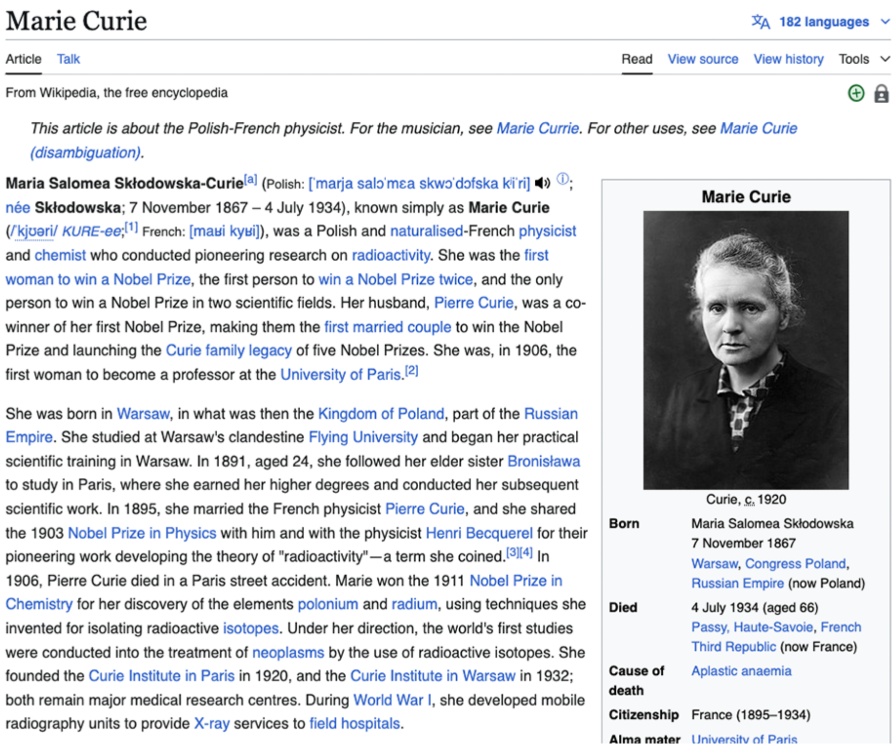

Figure 3.13.: Screenshot of the first fold of the Wikipedia article about Marie Curie. Taken from Michaelsexton2003 (2024a).

The Chrome Extension that sets the image description process in operation must first be integrated into the Chrome Browser. To do so, the developer mode must be enabled, because the software is not signed for the official marketplace and can only be used in that way. Thereupon, the algorithm can be executed by pressing a shortcut (Windows: Ctrl + Shift + 1, Mac: Ctrl + 1).

The result of the executed algorithm is exhibited in the next Figure 3.14. This is shown by the blue raycasting lines and the red dots representing intersections between rays and text elements. It can be seen, that the algorithm worked successfully, because the image description was appended to the original paragraph (see blue rectangle).

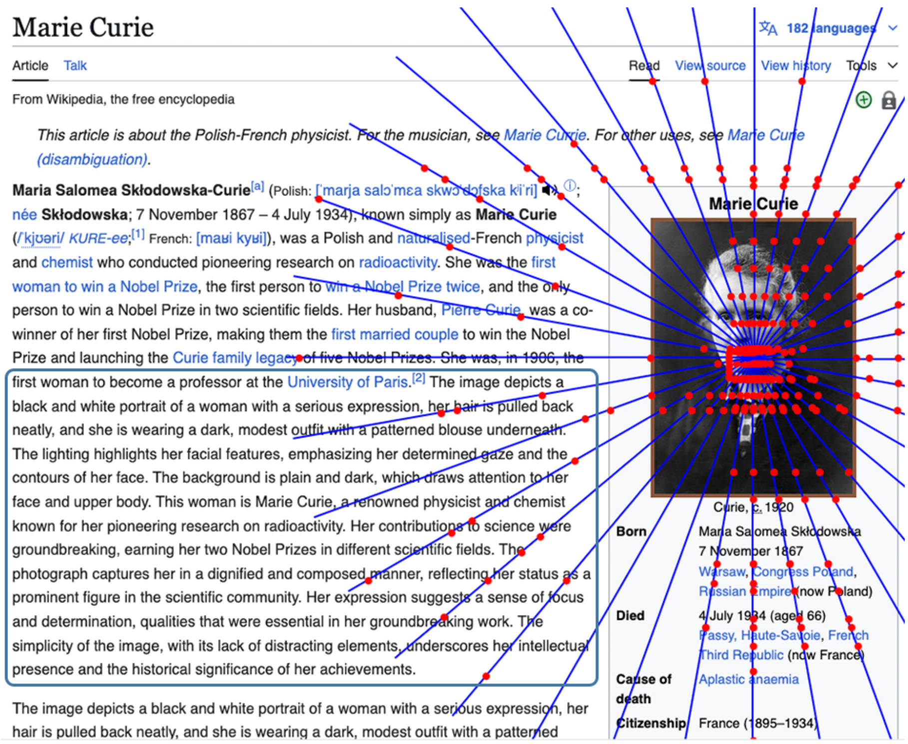

Figure 3.14.: Screenshot of the first fold of the Wikipedia article about Marie Curie. The final operating principle of the prototype is shown by raycasting, emitted in the center of the image, and the resulting generating image description within the blue rectangle. Taken from Michaelsexton2003 (2024a).

In order to illustrate the resulting image description better, it is shown in the following citation. First, the description includes features of the scientist's appearance, such as her hairstyle and facial expression. It also describes the photographic setting of the picture, i.e. that it is a black and white portrait.

Then her career is briefly discussed, followed by an explanation of how her appearance conveys her determination.

> *"[...]. Her husband, Pierre Curie, was a co-winner of her first Nobel Prize, making them the first married couple to win the Nobel Prize and launching the Curie family legacy of five Nobel Prizes. She was, in 1906, the first woman to become a professor at the University of Paris.[2]* ***The image depicts a black and white portrait of a woman with a serious expression, her hair is pulled back neatly, and she is wearing a dark, modest outfit with a patterned blouse underneath. The lighting highlights her facial features, emphasizing her determined gaze and the contours of her face. The background is plain and dark, which draws attention to her face and upper body. This woman is Marie Curie, a renowned physicist and chemist known for her pioneering research on radioactivity. Her contributions to science were groundbreaking, earning her two Nobel Prizes in different scientific fields. The photograph captures her in a dignified and composed manner, reflecting her status as a prominent figure in the scientific community. Her expression suggests a sense of focus and determination, qualities that were essential in her groundbreaking work. The simplicity of the image, with its lack of distracting elements, underscores her intellectual presence and the historical significance of her achievements.*** *" (OpenAI, 2024d)*

As required in Section 3.4, the Chrome Extension can operate on all operating systems, as long as the Chrome Browser is installed. Furthermore, users can simply use the shortcut shown above to set the algorithm in action, which extracts images and text from the webpage, describes the images based on them, and finally inserts the resulting text back onto the original webpage.

Regarding the non-functional requirements, it should be noted that a suitable performance has been achieved with available resources, which allows individual images to be processed within 5 to 10 seconds. The quality of the descriptions could be more precise and more context-sensitive, but this would require further research and development of the LLMs. However, it is very important to note that users should not use this software on sites that contain private or other sensitive data. This might have been possible with LLaMA-3:70b, since this LLM works offline, however, caution is currently recommended. Finally, the usability goal was achieved as requested. The image description process can be triggered by pressing a key and blind users are continuously informed about updates. In addition, users can continue navigating as usual while the software is processing data.

3.8. Summary

In conclusion, it can be said that it was only possible to implement the prototype with a few compromises. The biggest challenge was not to generate image descriptions as originally assumed, although this could also be greatly improved, but the organization of the images and texts from web pages. An ideal scenario would have been for the AI to handle the entire process independently and fully autonomously. However, this approach failed on the one hand due to the costs, and on the other hand, because of the decreasing reliability while the length of the prompts increased.

During the empirical development process, it was determined that the raycasting algorithm, together with the AssistantsAPI, was best suited for this prototype. The implementation of tha raycasting feature in the Chrome Extension as well as the development and deployment of an Express.js app, that communicated to OpenAI, allowed to verify the functionality of the prototype on Wikipedia articles.

Regarding the LLMs, they are only partially suitable for this project. They are excellent at creating a detailed description from little input, such as an image or short prompt. However, their weakness is that they inaccurately reproduce information provided by prompts, and apparently the length of the prompt increases the error curve. In addition to that, the

They can creating a detailed description from little input, such as an image or short prompt. However, some generated descriptions in this research could rather be considered as "stupid". This shows, that Generative AI is still not close to reach human intelligence, even though the generated natural language might seem intelligent. On top of that, they require powerful computers, such as a Shadow PC for the LLMs from Meta, in order to operate within a reasonable time frame. Generative AIs like the Assistants API should be used with caution during development, as each prompt consumes available budget, and iterative and empirical research requires many prompts with fine adjustments. In comparison to the AssistantsAPI, ChatGPT has predefined and unchangeable settings. Despite of that, ChatGPT could be used for this research without any deficits. Moreover, the quality of image descriptions from OpenAI's LLMs were better on average than those provided by Meta. Furthermore, the GPT-4o service is easier to integrate, because it does not require the explicit setting for either a text or vision model, instead GPT-4o detects the necessary settings by itself.

The next chapter will investigate whether the achieved results will also work in practice. To do this, interviews were conducted with experts, who tested and evaluate the prototype.

4. Evaluation

Based on the theoretical considerations, a prototype was developed with the aim of describing content-relevant details of images on websites and improving navigation efficiency for blind people. A qualitative study was conducted to determine whether the use of this prototype actually has a positive influence on the user experience of blind users or not. Therefore, expert interviews with blind people were carried out.

In order to gain a comprehensive insight into how blind people interact with images, to understand the strategies used by them to navigate webpages and to find out whether the prototype supports them in these processes or not, the following subordinate research questions were derived for the qualitative study:

1. Exploring which tools participants currently use in their daily lives to navigate websites and what barriers arise.
2. Elaborating participants' experiences in perceiving images on webpages and to which extend alternative texts are useful to them.
3. Identifying the degree to which the developed prototype is useful for the participants and their reasoning.
4. Discussing the affect that the participant's professional background has on the application of the prototype (e.g. usability).
5. Identifying improvements of the participants for further development of the prototype.
6. Exploring what kind of tools participants generally prefer to see invented in the future.

In order to address these questions, the next section will exhibit the entire procedure of the qualitative study, starting from selecting a research method, explaining guideline-based expert interviews, defining a target group, collecting the data through interviews, demonstrating the results, interpreting them and pointing out limitations of the research scope.

4.1. Method and Procedure

The qualitative research method was chosen to answer the research questions. This has the advantage of being able to interview people in more depth and to capture their individual attitudes holistically. In the following section, this qualitative scientific approach as well as the empirical study and the evaluation process are described.

4.1.1. Qualitative Research Method

The field of empirical research consists of two sorts of analysis methods. These are, on the one hand, quantitative research methods, and on the other hand, qualitative research methods (Gläser, 2006). Quantitative methods refer to figuring out causes and effects of situations, in order to determine regularities and forecast possible impacts (Merriam & Tisdell, 2015). They focus on certain characteristics of participants to elaborate the data and adjacently apply statistical methods for assessing them (Schaffer & Schaffer, 2019). In contrast to that, qualitative research brings individual persons as whole into focus. In this sense, certain aspects play an important role, such as the interpretation of their thoughts in specific situations, their view on the world and the meaning they attribute to their experiences (Merriam & Tisdell, 2015). In order to obtain these information from individuals, interviews or group discussions are conducted and analyzed using verbal evaluation methods (Gläser, 2006). It has to be mentioned, that there are many more different types of verbal evaluation methods (Merriam & Tisdell, 2015), thus the right one has to be picked for each research case.

The concept of the developed prototype is based on the findings of the theoretical research and the implementation relied on available resources such as time, budget, third party services and further on. Even though the development process was realized in all conscience, an in-depth evaluation is required, that shows to what extend the prototype is beneficial and user-friendly for individuals. The evaluation needs to involve practical experiences using screen readers, various navigation strategies on websites, individual preferences of which image features are considered as relevant to describe and personal desires about evolving technologies in the future. The most suitable way to process these outcomes was to utilize the qualitative research method, because it centers individuals.

4.1.2. Research Tool

In this study, interviews with experts were conducted with the assistance of a guideline. Therefore, it is particularly important to define the term "expert", that is commonly used for elitist people. Elitist refers to high-positioned people such as scientists, politicians etc., because they are attributed to have particular knowledge in a certain discipline. In addition to that, they are often expected to provide solutions for specific problems on demand. However, whether a person is an expert in the context of an expert interview strongly depends on the interest of the study. In this sense, an expert occupies the role as a "medium" and provides the researches with crucial information about the investigated social context (Gläser, 2009). As part of the present research, visually impaired and blind people were defined as the interview experts, because they use screen readers everyday and thus they face barriers that are of interest for this research. Hence, their experiences, thoughts, behavioral patterns for navigating on websites and personal opinions about image descriptions are of great significance.

As mentioned, the interviews were carried out using a guideline. A guideline's purpose is to assist the interviewer for finding out thoughts, interests, emotions and further characteristics of the expert, that are relevant for the study's objective. Besides that, a guideline also serves as a structure, red thread for the researcher and ensures that the statements of all interviewees can be compared. Thus the researcher can more easily maintain the conversation's direction and flow, that should add a natural atmosphere to it. In order to accomplish that, the interview questions should be expressed in everyday language and not read out as written, because reading out the questions accurately with a scientific choice of words would rather negatively impact the conversation and prevent obtaining in-depth findings. As a result, the usage of a guideline brings the collected data into a proper structure that leads to a better format for the qualitative analysis later on (Niebert & Gropengießer, 2014).

In order to create a guideline, it is expected that the researcher has already studied the respective topic in-depth from a theoretical point of view. This should ensure that the researcher has acquired fundamental knowledge that is necessary to formulate the appropriate questions for the given target group (Niebert & Gropengießer, 2014).

In the preparation of the guideline for this study, the methodology of Stigler and Reicher (2012) was used. The guideline is divided into a general and a detailed structure. The general structure consists of thematic blocks that are ordered chronologically. This is followed by the detailed structure, which consists of several subordinate questions. These are thematically linked within each block.

It is recommended that the first thematic block serves to create a suitable conversational atmosphere.

The guideline used for this investigation was thematically divided into five sections:

1. **Introduction:** Questions about demographics, job or education and age.
2. **Websites:** Questions about personal experiences with websites, used assistive technologies and thoughts about important image details.
3. **Prototype:** The phase where the prototype is demonstrated.
4. **Evaluation:** Questions determining the impact of the prototype and its usability.
5. **Future:** Questions about wishes for future developments.

4.1.3. Setup

According to Bortz (2009), planning and conducting qualitative expert interviews can be broken down into seven steps. According to Bortz (2009) these are described below:

1. Content-related preparation
2. Organizational preparation
3. Beginning of the conversation
4. Conducting and recording the interview
5. End of the interview
6. Adoption
7. Notes taken

In the first step, a guideline was prepared based on the findings from the theoretical and practical chapter (see previous section) as well as (Stigler & Reicher, 2012).

Simultaneously with the creation of the guideline, the process of finding interview experts began. To do so, involved people in this thesis, working colleagues, acquaintances and friends were contacted. As a result, they provided e-mail addresses and telephone numbers. One participant was directly found through the snowball system, by meeting another expert. Getting in touch with the experts by contacting them from person to person always worked, however, sending a mass-email to members of Odilieninstitut Graz was unsuccessful.

The individual meetings were arranged independently from each other, thus the location was sometimes even chosen spontaneously, whereby the participants were asked to propose a meeting point. They preferred to select places at the University of Graz, the Wall Central Library or Graz University of Technology

(Inffeldgasse). Due to possible weather changes and background noise the spots were mainly chosen indoor in learning areas or the cafeteria (at a less crowded times). Every participant received a declaration of consent (see appendix) in advance, stating that their data would be used anonymously and without the possibility of tracing them back by their provided information. In addition to that, they were asked to bring a private desktop device (a notebook), with the Chrome Browser installed. My own MacBook (16-inch, 2019) was also taken to the meetings for emergencies, in case setting up or using the prototype would fail on their devices. The MacBook was used for one interview. Two interviewees also brought braille display as additional equipment. Importantly, the prototype was uploaded on a private web server to make the download for the participants as easy as possible. Lastly, to completely ensure the functionality of the prototype during the interviews, a credit of ten euros was privately paid at the OpenAI billing platform to make the AssistantsAPI available.

Before the meeting started, we met in a public place, such as the main entrance or the staircase. In most cases, the blind person politely asked if they could put their hand on my arm, thus we could get to the interview location faster. During the route to the desired location and the first minutes after being settled there, a small talk was held in order to ease the atmosphere. At the point, when the interview started, the declaration of content was briefly repeated and it was pointed out that the conversation will be recorded with the smartphone and that the recordings stay anonymous. Subsequently, they signed the declaration without hesitation. Shortly before the conversation actually started, a quick recording test was done to double-check the recording quality and to avoid any loss of information. The smartphone was then placed in the middle of the table and pushed closer to the expert if necessary.

During the interviews, care was always taken to ensure, that the atmosphere remained relaxed, but simultaneously it was attempted to not drift too far from the topic. Nevertheless, longer excursions were sometimes unexpectedly useful, because they provided additional insights. In the course of this, it was sometimes also crucial to sensitively recognize the right moment, when an interviewee could be politely interrupted, in order to come to the follow-up question. One of the most successful methods for returning to the guideline was to take keywords from the following questions and integrate them into the context of the current answer. A central part of the interview was also that the experts test the prototype. For this, the verbal interviews were paused in order to move on to the practical part. The duration of the demonstrations varied depending on their computer performances, the speed of their screen readers and the possible installation difficulties. Every now and then they needed help to install the program. The participants had to open a few Wikipedia articles of their choice, execute the software several times and then return to the interview to start the evaluation of the tool and discuss future developments.

In order to provide a better overview, the exact process of the prototype test is listed below:

1. The participants were verbally guided to a website, where the download link for the software is provided. This website was explicitly created for this study.
2. They were given the task to search for one ore more arbitrary Wikipedia articles of their choice.
3. Then the participants had to press a keyboard shortcut to start the prototype.
4. Once the prototype has finished the image analysis, the participants were asked to take their time and read the article including the generated image descriptions.

The end of the interview was orally announced, while the recording device was switched off. Afterwards, there was still time to engage in another small talk and share personal information. In some cases, the content of the interview was even further discussed in an informal way. During these conversations, the interviewees received a small thank-you gift and some participants showed great interest of possible development progression. One participant encouraged to share the prototype, after the thesis had been handed in, in blind developers' platforms for free and open-source.

In the last step, after the experts had left the location, information such as time, location and additional notes were written down.

4.1.4. Participants

In the course of arranging the interviews, experts had to be found. Therefore, they had to fulfill two criteria and had to agree to be interviewed. These criteria were, that they were blind and were capable of navigating on websites using screen readers with optional further assistive devices. Other characteristics such as age, gender or profession did not play a role during the selection process.

A total amount of five persons (female = 1, male = 4) were interviewed for this qualitative research (*I 1* to *I 5*). All interviews happened in a time period from the end of September until the middle of October 2024. Conducting the interviews online would have been possible, however it could have lead to additional technical issues if they required human assistance. In fact, they were all conducted in person.

As exhibited in Table 4.1 the ages were situated between 21 and 39 years. While *I 1* to *I 4* had a bit of remaining visual perception, such as brightness, colors or even simple shapes, *I 5* was absolutely blind. *I 1*, *I 2* and *I 3* suffered from retinal impairments, whereas *I 4* and *I 5* had a defective optic nerve. All participants were unable to use their vision in order to access websites, consequently all had the same prerequisites. Every interviewee attended a university, however their study programs (e.g. Software Engineering or Translation) were different.

I	Age	Job/Study	Type of Visual Impairment	Vision
I 1	21	currently studying mathematics	Leber Congenital Amaurosis or EORD (Early onset retinal dystrophy)	only central field of vision, which is blurry; visual field of approx. 3 degrees
I 2	28	currently studying teacher education (specialization in inclusive education, philosophy and psychology)	Retinopathy	only light and dark perception
I 3	34	currently studying teacher education (specialization in inclusive education, German); works as part-time freelancer at the university	Retinal Detachment	remaining sight of 2 percent, only able to see outlines
I 4	22	currently studying Software Engineering; works as an IT-administrator at the university	Optic Atrophy (high fever during pregnancy)	a bit of colors, light and outlines, but no depth perception
I 5	39	currently studying translation (Italian and Spanish)	defective optic nerve	no vision

Table 4.1.: Overview of participants (*I 1* to *I 5*).

4.1.5. Transcription

Since the recorded data from the interviews were stored as audio files, they had to be transcribed to enable the analysis in the course of this work. For a proper transcription, a number of rules must be observed, which the researcher must follow consistently. According to Gläser (2006), there are no generic rules existent yet, which leads researchers to define and precisely document their own rule sets. The defined transcription rules, that were used in this thesis, are presented below:

- The interviewer is referred to as *R* (= researcher), while the interviewees are referred to as *I n* (= interviewee and number).
- The language spoken during the interview is retained in the transcription. Dialects are not adopted and sounds like "ahm", "hm" or "well" are removed, but repeated words are kept. Incomplete sentences are ended with "(...)" and incomprehensible passages are marked accordingly. Stuttering and filler words are only transferred if they are relevant for the content. Emphasis within sentences is marked accordingly, such as laughing.
- All places, names, institutes and other options that could enable the interviewee to be identified are replaced by placeholders.

The transcription process was manually realized in order to ensure data protection of the interviewed persons.

4.1.6. Evaluation Method: Qualitative Content Analysis by Mayring

The purpose of a qualitative content analysis is to analyze collected material that originates from a particular form of communication (Mayring, 2015), in this case the communication form are interviews. Mayring (2015) defined six essential criteria for the context analysis:

1. The analysis is based on communication, which includes various forms such as language, music, images, etc.
2. In order to conduct the content analysis, a fixed protocol of communication (of any kind) is required (e.g. a transcript).
3. The content analysis has a systematic approach to avoid free interpretation or impressionistic reading of the material being analyzed by the researchers.
4. The systematic procedure should be based on precise rules that also help others to review and understand the analyzed content.
5. An important aspect is that the systematic approach is theory-driven. Therefore, a text should be interpreted on the basis of a question derived from theoretical research. Furthermore, the analytical steps are also shaped by theoretical considerations.
6. The content analysis is a deductive method because it aims to reach conclusions about certain aspects of communication on the basis of assertions.

The analysis procedure according to Mayring and Brunner (2013) is characterized by the fact that a category system is applied for the content analysis. This means that the data corpus is not only analyzed as a whole, but is also divided into smaller parts. These parts are then grouped into categories based on content-analytical rules. By following these rules, other researchers should come to the the same results with the same data in the categorization process.

The categories for the category system can be defined in two different paths. These paths are deductive and inductive, meaning that deductive categories are defined based on the literature research, while inductive categories emerge from the analyzed content (Mayring, 2015). In the present content analysis, the categories were formed almost exclusively deductively. Only one subcategory (SC 2.2, see Figure 4.1) was formed inductively on the basis of the transcribed interviews.

Accordingly, three different forms of interpretation are distinguished for the qualitative content analysis. These are the Summary, the Explication and the Content Structuring. In the *Summary*, the collected material is reduced in such a way that the essential content is retained, but the data becomes more comprehensible. The *Explication* serves to add additional information to unclear text sections (terms, sentences, etc.) in order to provide a clearer understanding of the text passage. The aim of *Content Structuring* is to extract certain information from the material according to predefined criteria and to assess the material on that basis (Mayring, 2015). For the present research work, the Content Structuring was used to evaluate the interviews. This form of interpretation was chosen because it is best suited to present and interpret the experiences and perceptions of the blind interviewees. Therefore, *Content Structuring* is presented in more detail in the following Section.

4.1.6.1. Content Structuring

During the Content Structuring process, topics, content, etc. are filtered out of the material and summarized. In doing so, categories, which are formed based on the elaborated theory, determine which information should be extracted. Before this can be done, it must be determined in advance which source material will be used for the evaluation (Mayring, 2015). The evaluation material presented here consists of five transcribed interviews.

Mayring (2015) describes ten steps towards structuring the content:

1. **Determination of the analysis units:** To ensure a precise content analysis, three types of analysis units are distinguished. These are the coding unit, the context unit and the evaluation unit. The coding unit is the smallest part of the data to be analyzed (word), whereas the context unit represents the largest possible part of the data that can be assigned to a category (whole answer to a question). The evaluation unit ultimately defines the order of the data evaluation. In this research, the evaluation unit was the analysis of the five transcribed interviews. In doing so, the first category

was analyzed in each interview, then the second one, and so on for each interview.

2. **Theory-driven definition of the main content categories:** The main categories of the category system are formed deductively, so they were derived from the theoretical part (Mayring, 2015). In this study, five main categories were formed on the basis of the interview guide.
3. **Theory-driven definition of subcategories:** In the course of this step, the subcategories are formed, from which an entire category system follows. In this research 13 subcategories were formed (see Figure 4.1).
4. **Creating definitions, anchor examples and coding rules for the categories:** Definitions are formulated for the individual categories for the categories system. That determines which text passages must be assigned to a category (see fourth column in Figure 4.1). The anchor examples serve as particularly striking examples from the data and the coding rules serve to precisely define the categories. The anchor examples and coding rules can be seen in Figure 4.1.
5. **Material Analysis:** At this point, all text data is passed through and the corresponding text passages are selected and assigned for all categories (Mayring, 2015). In this research project, the interviews were printed out, examined in parallel, and the text passages were assigned to the category system using different colored highlighters.
6. **Search and extraction of results:** In this step, the selected data were extracted from the data corpus into an Microsoft Excel file.
7. **Iterative revision:** The defined categories in step three can be adjusted again, if necessary. The first pass can be seen as a test run.
8. **Paraphrasing:** Firstly, selected coding units are reformulated so that only the relevant content remains. Secondly, these are then paraphrased, leaving only the relevant text passages. Finally, the reformulated text passages are reduced so that redundantly coded units are removed.
9. **Summary of sub categories:** A summary of the subcategories of the main categories is carried out.
10. **Summary by category:** In this step, the contents of the main categories are summarized.

The results of the categories and subcategories will be presented in the next section.

4.2. Results

This section presents the results of the qualitative study. The presentation of the results is structured according to the categories (see Figure 4.1). The first

Category	SC-Nb.	Subcategory	Definition of category	Anchor examples	Coding rules
1. Personal Information	SC 1.1	Age	Records the age of the interviewee.	"Yes, I am 39 years old [...]" (I 5, 00:00:25-1)	How old is the interviewee?
	SC 1.2	Job/Study	Records which job/study the interviewee is currently doing.	*"[...] I am studying for a bachelor's degree in teaching and work part-time as a freelancer at the university."* (I 3, 00:00:39-5).	What is the person currently doing?
	SC 1.3	Type of visual impairment	Records thevisual impairment of the interviewee.	*"Yes, I have had a retinal detachment since childhood, which means that I have about two percent of my original vision left. [...]"* (I 3, 00:01:30-0).	Which visual impairment does the interviewee have?
	SC 1.4	Vision	Records how much the interviewee can still see.	*"No, nothing. It's defective. Like when a cable is defective in a computer, so you can't use it. So to speak, the interface between the eyes and the brain, the optic nerve is deformed and yes."* (I 5, 00:01:04-6)	How much can the interviewee still see?
2. Navigation on websites	SC 2.1	Devices and tools	Records which devices and tools the interviewee uses.	*"JAWS with braille display and sometimes with audio output. [...]"* (I 1, 00:02:40-7)	What devices and tools does the interviewee use?
	SC 2.2	Use of AI	Records how interviewees utilize AI tools.	*"Yes, for example I can let BeMyEyes describe images on Social Media for me. [...]"* (I 2, 00:06:36-2)	How does the interviewee currently use AI?
	SC 2.3	Major problems	Records the main problems interviewees face while navigating.	*"[...], when certain elements cannot be navigated by using the keyboard"* (I 2, 00:03:22)	What are the main problems the interviewee has when navigating websites?
	SC 2.4	Problem-solving strategies	Records the strategies that interviewees use to solve problems while navigating.	*"The easiest way ifor returning to the start is using the browser history"* (I 3, 00:13:17)	Which problem-solving strategies does the interviewee use when navigating websites?
	SC 2.5	Experience with visual content	Records the interviewee's previous experience of accessing visual content on websites.	*"A picture contains a thousand details, so I would just need the relevant one."* (I 3, 00:07:43-0)	Which experiences has the interviewee had with accessing visual content on websites?
3. Evaluation	SC 3.1	Continued use of the prototype	Records whether the interviewee would continue to use the tested prototype and the reasons for or against using it.	*"However, actually, I would use something like this for a website with a large amount of images. Like Wikipedia or some cooking recipe, [...]"* (I 4, 00:00:21)	Would the interviewee continue to use the prototype? If so, why? If no, why not?
	SC 3.2	Wishes for change	Records further wishes from the interviewee to be implemented in the prototype.	*"So, you can only do that with configuration options. Do I need the scenery here, do I need the context of the image here, what does that say and so on."* (I 5, 00:01:46-3)	What else does the interviewee want from this prototype? What suggestions for improvement does the interviewee have?
4. Future	SC 4.1	Future usage of the protoype	Records what the interviewee suggests how the prototype could be furtherly applied.	"Otherwise, of course, if it's faster, it's definitely a good way to make such things more accessible." (I 1, 00:03:24-0)	What does the interviewee suggest in which context the prototype could be used in the future?
	SC 4.2	Other future tools	Records what the interviewee thinks could be future tools that make images accessible and improve navigation overall.	*"I would find AI useful in a context that describes stuff that I want to know, but simulatenously. If I were to use some glasses or some camera."* (I 4, 00:07:48)	What does the interviewee think could be future tools that make images accessible and improve navigation overall?

Figure 4.1.: Categories and subcategories (SC) of the interviews.

category is not included, since it has already been demonstrated in Section 4.1.4.

These categories are:

1. Navigation on Websites
2. Evaluation
3. Future

4.2.1. Navigation on Websites

All interviewees were asked which devices they use for website navigation, if they use AI, what the biggest hurdles are for them while browsing the internet, which strategies they use to overcome these hurdles and how they perceive visual media on webpages. The results are demonstrated within the next Sections.

4.2.1.1. Devices and Tools

For the navigation on websites all interviewees used smartphones as well as desktop devices (notebooks). Everyone used an iPhone as their mobile device, and *I 4* used an Android phone as a secondary device. In terms of desktop devices, either Windows (*I 1*, *I 3*, *I 4* and *I 5*) or Macintosh (*I 2*) were used. The notebooks were each used with screen readers and, when needed, with a portable braille display. For Apple devices, both smartphones and notebooks, VoiceOver was used as the screen reader software, which is integrated by default in Apple devices. The participants who owned a Windows notebook mostly used NVDA, however *I 1* stated to use JAWS. In contrast to that, *I 4* explained to be denying JAWS for the following reason:

> *"[...] JAWS is expensive and I'm anti-establishment insurance companies that try to profit from people with disabilities from tools that they need on a daily basis. This is firstly a moral objection. What I have in secondary NVDA is open source compared to JAWS, which is proprietary software, and it's much more customizable [...]"* (I 4, 00:07:45)

In addition to that, *I 4* also mentioned to be a certified NVDA user and chose this screen reader because it is free of charge and has great community support.

4.2.1.2. Use of AI

The second question in this Section focused on the current use of AI products by the experts. *I 2*, *I 3* and *I 4* reported that they either use AI tools or at least had experiences with them in the past. In this sense, *I 2* outlined the application of the app *Be My Eyes* on social media, where pictures can be described on demand. The second interview person (*I 2*) also mentioned to use this app in other scenarios, such as in public transport to capture photos out of the window and get a description of the environment.

Be My Eyes turned out to be an app, that might be used by many blind people. *I 3* mentioned in the interview, that the app was originally used with human assistance, thus blind or visually impaired people could take a photo and ask sighted people via the app to describe the photo for them. Nowadays, the app uses an AI for analyzing the pictures in addition to the human assistance.

4.2.1.3. Major Problems

The interviewees reported various hurdles that they encountered on websites. As an example, *I 1* described that it took practice and sometimes longer to arrive at the main content of a webpage using a screen reader. A reason for this is, that some other elements (e.g. texts, buttons, etc.), that are no relevant for the search, have to be conquered on the way. Regarding this, *I 2* added that keyboard controls are essential for screen readers, however it is not always guaranteed that elements are reachable with the keyboard. Thereon, *I 5* complained, that advertisements on webpages often severely impair the reading flow, because they are "in the middle of the content" and thus disrupting the screen reader. In order to encounter this problem, *I 5* recommended to use most recent ad-blockers. Referring to this issue, *I 3* added that their are platforms, e.g. for games, that show advertisements, however the developers do not have control of the advertisement itself, thus they could be inaccessible for screen readers. While these advertisements are shown, the user has to wait until a close button appears. However, closing these windows through a screen reader could be impossible as well. Other obstacles are caused by e.g. cookie-banners or CAPTCHAs. Cookie-banners sometimes require the screen reader user to scroll down to the bottom of the webpage in order to accept them (to close them). In other cases, cookie-banners overlap the screen, so the blind person has to patiently try to find the close button. Regarding CAPTCHAs, *I 5* provided an example that can have severe causes for blind people. For instance, it is assumed that a new online platform was developed where people could order food. In case this platform had a CAPTCHA right before the final submit of the order and this CAPTCHA was inaccessible for screen readers, thus hundreds of blind people could not receive their food.

Also, *I 3* stated, that many issues occur, because accessibility is not sufficiently taken into account when new updates were rolled out, where important screen reader features were no longer available. The fourth interviewee (*I 4*) claimed, who has a software development background, that hurdles are also often caused by websites, that run on JavaScript. The interviewee reasoned that this type of problem arose because the developers of JavaScript websites have to explicitly implement the accessibility of the elements, but they often forget to do so. This can cause blind people to neither being able to reach certain elements on the webpage, nor even knowing that they exist. The interviewee suggested a solution by rather focusing on native HTML elements as a developer, because this benefits the screen readers better. Lastly, *I 5*, who had a language translation background, stated the following problem:

> *"Let's assume an American data company creates a homepage and then has it localized, as they say in translation technology. The speech, i.e. the HTML code, is not synchronized or localized in German. Then you have the problem that you get a German website read out in English."* (I 5, 00:05:20-5)

4.2.1.4. Problem-Solving Strategies

The interviewees mentioned different ways of navigating efficiently through websites and strategies for overcoming hurdles. *I 1* explained that touchscreen navigation works very efficiently on the iPhone. Therefore, the blind user swipes their finger over the web elements and these are immediately read out via voice output. Furthermore, *I 2* said that sometimes directly navigating to the shopping cart is not possible, so the address of the shopping cart must be entered in the address bar. However, in order to do so, a sighted person has to provide the address of the shopping cart when visiting a website for the first time. This interviewee also mentioned, that the navigation efficiency was hugely increased by jumping from heading to heading or from link to link. In particular, headings made it easier for *I 2* to get an overview of the web content and to subsequently delve into certain text passages. *I 3* explained that software on Windows had some customization features, which were pretty helpful. As an example, these involve listing all links or text fields at once.

Another question was how the participants can recover, if they get lost on a webpage. Regarding this, *I 3* mentioned to go to the page start for another attempt. Therefore, the browsing history is used to find the latest visited website. Referring to getting lost, *I 3* also described how the app Be My Eyes with the recent AI features could help to solve CAPTCHAs by describing them. *I 3* mentioned that there is no guarantee that Be My Eyes can aid every time (*I 3*).

Furthermore, *I 5* complained, that bad language settings would ruin the audio output for screen readers, so the suggested solution for this was to use the braille display instead, so the user can directly access the braille signs and interpret by themselves.

Generally, in order to cope with the most situations that were mentioned above, blind users had to use screen readers with different and personalized strategies. However, in fact, *I 2* explained that websites could be implemented in a more accessible way. Therefore, developers would have to follow the standards of the web content accessibility guidelines. Following these rules during design and development, not only blind users, but also people with various other disabilities would benefit from it.

4.2.1.5. Experiences with Visual Content

Perceiving images is a major issue for blind people on the web and also the central topic of this thesis. All interviewees regularly faced similar burdens: missing or badly expressed alternative texts. They also had in common, that mostly only certain image features were relevant for them, however it exactly depended on their personal preferences.

I 4 described a Google tool from one to two years ago. With this tool, images could be sent to Google via right-click, these were interpreted there and then made available to users. However, this tool was not found to be very helpful because it only answered e.g. with one word. *I 4* never really used this tool, because why would the person need to know e.g. if there is a photo of table without any further information.

Interviewees *I 1* and *I 2* complained about unusable image descriptions. In this sense, they can be unusable by either having no alternative text at all (or they contain the picture's file name), or they are formulated in a bad way (*I 2*). According to *I 1*, an image description that only described "picture with two persons" is pretty much useless. Thereon, *I 5* generally explained the following:

> *"You have to distinguish between those who are blind from birth – I don't know if you've noticed that yet – and those who have gone blind later on. I think people who have gone blind later on will still want to know what this picture is about."* (I 5, 00:06:41-1)

Regarding the details of an image, *I 3* clarified that it is hard to generalize which features are important for individuals, because an image consists of thousands of features. However, *I 3* also mentioned that only the relevant image details for a specific context would be enough to be captured —the context is essential.

When the interview person was asked to tell important features of an example, where persons are at an honorary doctorate award ceremony, the interviewee replied:

> *"What they do or say, in particular, is important now. Not how they look or whether they are wearing a hat or whatever."* (I 5, 00:07:32-1)

In contrast, to give another example, *I 3* said when shopping online, it was particularly important to see what the products looked like. However, *I 5* mentioned that in these or similar situations, it was not only important for alternative texts to describe the image features, but also to help blind users figure out, if a focused button is the correct one and which action will follow. In case a button was not labeled correctly, the screen reader only announced "button", so a button with only a shopping cart icon is not enough.

Some participants told about experiences, either from the the past or from the present, when they were using AI tools to capture images. *I 4* talked about tools that had already been tested in the past. These included an integrated function of the Chrome Browser that was supposed to interpret images. This reportedly hardly worked, but the interviewee was confident that today's apps and AI capabilities would overcome these difficulties. This is expressed again in the following quote:

> *"If I were to take a photo of you it would say 'A photo of a man'. Not really detailed but now if I take it with the emergence of ChatGPT with GPT models [...]"* (I 4, 00-16-47)

In addition to that, *I 2* reported a few experiences with modern AI tools. The interviewee was using AI since the year before and was totally fascinated about its features and capabilities (*I 2*). Therefore, the interview person mentioned:

> *"Yes, for example on social media I can have the pictures described via Be My Eyes. [...]"* (I 2, 00:06:36-2)

I 2 did not only use Be My Eyes for social media purposes, but also in other daily situations. The interviewee mentioned that they sometimes take photos out of the window during the bus ride and then the AI describes them.

I 5 mentioned another serious problem. When it comes to verification options on a website, to determine whether a person is actually a person and not a robot, problems can occur. As an example, when blind users want to use online banking and they face such a verification method, which are originally designed for sighted people, and audio captures are not available, they cannot proceed.

However, it could be concluded from all five interviewed persons, that all preferred information from images, that is related to the webpage's context.

4.2.2. Evaluation

The following subcategories refer to the interview phase that took place after the prototype had been tested. During this phase, the interviewees were asked whether they would continue to use the prototype and what further developments they would find useful.

4.2.2.1. Continued Use of the Prototype

All interviewees agreed that they would continue using the tool or continue testing it extensively. At first, *I 4* had problems using the tool, however after the functionality became clear, the interviewee agreed to be interested for further testing. All other interviewees had no difficulties using the tool.

The quality of the generated image descriptions, and the connection to the webpage's context seemed adequate for all interviewees. *I 1*, *I 2*, *I 3* and *I 5* explicitly mentioned, that the functionality would be rather useful for a large amount of text with many images. As possible application fields, *I 4* mentioned that using this prototype would be great for pages such as Wikipedia or recipes, *I 2* only referred to Wikipedia. *I 3* and *I 5* meant that using this prototype would be beneficial for large PDF files, whereby *I 3* said that the embedded image description into the text was not necessary. In the following a statement by *I 2* after the testing phase is shown, that emphasizes the importance of the prototype:

> *"Wikipedia contains a lot of pictures. And there are also a lot of pictures that are very meaningful. Because now, for example, with the diagram and all that, without the tool I would have had no information at all."* (I 2, 00:00:24-6)

4.2.2.2. Wishes for Change

I 1 mentioned, that the tool was generally good, however a wish got expressed to increase the generation performance of the tool. The second and fourth interviewee were curious to examine the prototype on other websites as well, to determine whether it works just as well as on Wikipedia.

I 4 and *I 5* had pretty similar wishes about further functions of the prototype. It should have more interactive possibilities to fine-tune the results. This could occur via a chat box, where further questions could be asked (*I 4*), or as *I 5* expressed, there could be a configuration option to decide, how much influence this tool could actually have on webpages, and to select if the context or the scenery of an image would be required.

I 3 wished to change the AI formulations in a way, that they sounded more natural. The interviewee complained, that they currently sounded too machine-like. The person added an interesting statement:

> *"As if it were somehow trying to sell me the picture or something."* (I 3, 00:02:18-7)

The main concern of *I 2* was how the data would be treated by the AI provider. However, participant *I 3* seemed to accept the state-of-the-art of AI and expressed hope, that this would improve in the future. The reason was, without using AI there would probably be more often no description at all. Moreover, *I 2* talked about improving the generation time performance. If the image analysis process took too long, the images could be described in the background, so blind users could continue navigating, but immediately access images if they were ready. The participant desired that the prototype jumps through the images once all have been prepared. *I 1* asked whether this prototype could also be used on smartphones, as far as the Chrome Browser was installed.

4.2.3. Future

In the following, the last category will be evaluated, which includes possibilities and wishes for the future. The focus here is primarily on a possible future development of the prototype, but also on general innovative solutions that do not yet exist.

4.2.3.1. Future Usage of the Prototype

The five interviewees thought that this prototype was usable and explained various application scenarios. All the interviewees agreed that the prototype was best suited for long articles or other digital products with a lot of text and images.

I 1, *I 2* and *I 4* would use the prototype for long web pages such as Wikipedia articles, recipes, etc. *I 3* emphasized that the usability of the prototype would depend very much on the field of application. The interviewee described that the embedded image description could be useful to allow image descriptions to be copied out of the document, since standard alternative texts cannot be copied with the screen reader.

I 5 described that a complete image description for common research tasks at the university would take far too long, so you would have to narrow it down. *I 2* and *I 4* stated that they had to test the prototype on several websites in order to provide more accurate feedback on this question.

4.2.3.2. Other Future Tools

In the last question, the interviewees reported various ways of dealing with a lack of image accessibility. On the one hand, *I 2* talked about the lack of standards for suitable image descriptions and about descriptions that are still very subjective, but also about problems that could be more effectively solved on the development side (*I 2*).

I 4 was very clear about the problem of the lack of accessibility awareness on the part of developers. In this regard, *I 3* stated that some new innovation would not be necessary at all if developers would just pay attention to accessibility measures. Another problem is that many authors of image descriptions are simply not sure what exactly they should write in the image description or they are not able to describe a complex graphic at all. *I 3* thought that in such cases, AI could be very helpful because it can do it easily.

According to *I 3*, the needs of blind people with regard to images remain individual. Therefore, in the future, there could be functions that provide different details in any level of detail. This level of detail could be determined by the user for each image feature. This way, blind people would not have to listen to image descriptions for minutes and could navigate more efficiently. In this context, *I 5* delivered the following recommendation:

> *"I don't think it will be down to one tool. Rather, you will always need a certain set of tools and, of course, the appropriate training to be able to use such things. [...] And that is why it is important that such tools are developed with people, by involving them in the development, not in programming, but in the projects, in the responsibility. [...]"* (I 5, 00:13:04-9)

Regarding the current capabilities of AI, *I 3* criticized the way how AI language is formulated. The participant claimed that it would mostly be expressed in a way, thus it sounds like AI. Another crucial aspect for *I 3* was, how the images would have been described if the images would not comply with American's values, since the OpenAI is a company from the US. The question is if they were still being described or only described in a specific way.

The first interviewee, who was studying mathematics, wanted a screen reader with better control options and better accessibility of multi-text formulas. Furthermore, *I 2* was overall rather satisfied with the state of the art, but concerns about privacy and security were still reported.

I 2 mentioned several times that they liked to use smartphone apps for various tasks. *I 3* suggested a better AI option for making handwritten markings, texts, etc. in PDFs more accessible for screen readers. The interviewee with a software development background preferred tools with many interaction options. Furthermore, the person mentioned that other tools such as smart classes or camera tools would also be desirable. However, *I 4* would not change the screen reader itself because of the existing habit.

4.3. Discussion

According to Nielsen and Pernice (2010) sighted people can recognize images in their peripheral field of vision in a fraction of a second and can also decide for themselves which image details they want to look at more closely. In contrast, since images require sight to be perceived, blind people are strongly dependent on alternative texts and the accuracy of their provided information (Firth, 2024). This was in line with what the interviewees reported in the study. They mentioned that it is still often difficult for them to get the information they really need from images on websites. One reason is that people are individuals and have different interests. They also agreed, that the situation is further complicated by the fact that they are dependent on how either the developers of a website write the alternative texts or how AIs describe the images. Levc (2024) also stated:

> *"[...] But with the image descriptions or with the alternative texts, which are really created actively, so to speak, it is often really just the decision of the person creating the alternative text, based on what they think is important. [...]"* (Levc, 2024)

Since 2015 the app BeMyEyes has been available which all study participants gave very positive feedback on. With the help of this app, blind people can have photos or live videos described by sighted people. In August 2023, BeMyAI was released based on GPT-4 for beta testers that can generate image descriptions with AI. The focus of BeMyEyes and BeMyAI is on describing the blind user's current environment and is therefore offered as a smartphone app. However, the focus of this master's thesis is on images from websites, which are contextually described using AI and where the resulting descriptions are attached in the main text. The people interviewed were not yet aware of such a tool. Only one interviewee mentioned a tool that was integrated into the Chrome Browser and described images with a single click. However, the interviewee was dissatisfied with this tool because it described a table, for example, only as a "table" and provided no further information.

The evaluation showed that the study participants found the prototype helpful. As mentioned by Foster (2023) recent Large Language Models utilize der Transformer architecture, that improves the context awareness and understanding of provided text data by the user. This technology was fine-tuned in the course of the prototype development by using the AssistantsAPI, in order to reduce redundancies in the prompts to keep the costs low (see Section 3.5.3.3). The blind people rated the AI-generated image descriptions as accurate and confirmed that the contextual nature of the tool added value for them. In addition, the participants mentioned several scenarios in which this prototype could be used, such as PDFs, recipes, etc. In general, they thought that the way images are described and the image description embedded in the body text is useful for very long documents with many images. Since nowadays LLMs are already able to process multi-modal data apart from text, such as images, videos or audio, there could be futuristic solutions that even go beyond the mentioned examples (Kulkarni et al., 2023). Furthermore, all study participants were able to use the prototype quickly and easily because it only needed to be integrated as a Chrome Extension and only required a shortcut to analyze the images on a web page. Ensuring that the software is so intuitive to use was the requirement in the planning of this prototype (see Section 3.4).

Firth (2024) stated that the accessibility tree modifies the DOM tree in a way thus screen readers can directly access the relevant HTML elements to read out loud. This is relevant for blind people in order to be included to all information on a webpage and not to become lost in complex layout structures. These accessibility measures were combined with the elaborated usability principles in Section 2.1.2.2 and implemented using the ARIA-tags in Section 3.6.4. As a result, the constructed auditive user interface, that notifies screen reader users once a new image has been described, seemed intuitive enough that the participants could immediately find the image description into the main text of the Wikipedia articles.

The interviewees also expressed criticisms and made suggestions for improvement. One participant found that embedding images in the text was not necessary for them to be able to see images, but that this function would make it easier to copy the image description to the clipboard. One interviewee claimed that the analysis time of the images by the AI should be shorter. This is currently not possible without losing image description quality (see Section 2.2.3). Two other interviewees expressed the wish that future developments of the prototype would include configuration options that would allow the user to adapt the functionality of the tool at any time. This goes in line with Stangl et al. (2020) that blind people have different interests in which image features the prefer to perceive. In addition, the image descriptions were criticized by one person because of the AI language, as they sounded too "artificial" on the one hand and on the other hand, an American bias could be behind them, since OpenAI is an American company. Furthermore, wishes were made that the program would be applicable to other websites. In Section 3.3 it was stated that the prototype theoretically works on other websites as well, however Wikipedia was only chosen to reduce complexity. Furthermore, the data security of OpenAI should be more transparent. Derner and Batistič (2023) also claimed that there are certain risks regarding the used data by companies like OpenAI. This tools are used by the global public and there is a lack of transparency how the data are treated. According to Kosinski and Forrest (2024) criminals could use various attacks to work around security measure of the Large Language Models and leak personal or sensitive data.

5. Lessons Learned

This chapter serves to reflect on experiences and successes, impasses, approaches and the findings of the course of the master's thesis. To do this, the chapter is divided into three parts: literature research, design and development and the evaluation of the prototype.

5.1. Literature Research

In the course of the theoretical research for this master's thesis, the topic of blindness was addressed first. The results of this chapter showed that visual impairment and blindness have neither a simple nor a globally uniform definition. Furthermore, it was found that most people referred to as blind have some residual vision, which is, however, insufficient to perform everyday tasks.

Various methods were discussed how blind people can use assistive technologies to navigate on websites. Blind users are just as sighted users able to skim websites. The difference is, that sighted people do this visually and blind people need screen readers with audio output or braille displays. However, screen reader users repeatedly encounter barriers because websites are either not developed in a sufficiently accessible way or elements that can be captured by screen readers are not sufficiently described.

Blind users can also access images through audio or their braille display. This is only possible if an alternative text is provided. Whether developers forget to add this alternative text or write an unusable formulation, the information becomes inaccessible for blind users. There are currently no standards for writing image descriptions can be created, so it depends on the individual developers which information they include. Sighted people can look at the image details and include those they consider relevant. For blind people, this is not so easy, because the alternative texts must contain the relevant information, but what is relevant for people depends heavily on many factors, such as personal interests, the context of the internet research, etc. In case no usable alt text exists, blind people are dependent on sighted people to get a picture described.

The last part of the literature research dealt with the topic of Generative AI and its potential applications to describe images automatically and meaningfully.

It was found that many different types of AI and large language models exist, which have various differences in terms of performance, costs, quality and security. A particular insight was that natural language processing is not a new concept, however nowadays technical resources, such as large amounts of data and powerful machines, could make it possible to generate meaningful content. In addition to that, the discipline prompt engineering provides strategies and patterns for users to maximize the output quality.

5.2. Design & Development

During the empirical development of the prototype, two types of challenges arose: extracting content (images and text) from web pages and dealing with Generative AI in terms of quality, performance and costs. Both components were necessary to make the prototype operational, but this required two separate empirical investigations.

The main finding when extracting data is that it is currently still very difficult to access the DOM automatically and to identify specific data. With previous solutions, user interaction with the DOM-tree is usually necessary or the parser is optimized for certain websites. Even though the AI was able to capture data from the HTML code and screenshots, the results were unsuitable for this project due to hallucinations in the content as well as high costs. In contrast, ray-casting was a better solution for extracting images and texts from the webpage. This technique is normally used in games, however the operation principle showed potential in this development, because it can be used to, firstly, detect neighboring objects on the webpage, and secondly, parse the content without mistakes.

With regard to the generated image descriptions, it can be said that Generative AI already shows great potential. Generally, the more details about an image are provided within a prompt, the more precisely the AI can produce the desired image description. In some cases, it leads to better results when prompts and responses for a single image are split into a multi message conversation. Thereon, various tests with the AI showed that it is possible to generate an array, which can immediately be reused in the project's code, instead of receiving a randomly formulated answer in natural language. The costs have to be taken into account, because they can quickly accumulate when the communication with the AI contains a lot of content. It is contradictory that the prompts of this developed prototype have to contain as much information as possible, but have to simultaneously be as minimized as possible to reduce costs. The Assistants API worked very well for this.

Furthermore, possible workarounds for the development process had to be found for the AI research. The reasons were that the required computing performance was only available through external services, otherwise private would have risen rapidly. Finally, it also meant a greater effort to provide the prototype with an infrastructure that allows the tool to be used portably on other devices. To do this, further aspects such as server costs, encryption, etc. had to be considered. Since it was prohibited for the Shadow PC to be used as a server, the final prototype was implemented only using OpenAI's services.

5.3. Prototype Evaluation

One of the most valuable experiences of this work was getting in touch with visually impaired and blind people. It was interesting to see how they use assistive technologies to navigate on websites and overcome barriers, because this provides important information to develop software for this target group. In addition to that, the contact strengthens the empathy, sensitivity and gives a broader understanding of inclusion. This can help to break down prejudices and enables one to change their own perspective, which has a lasting effect on the way to develop websites and write alternative texts in the future.

The results of the qualitative study were positive. The interviewees rated the prototype as helpful, but they suggested configuration options so that the generated image descriptions could be adapted to personal interests. However, all the interview persons agreed that this prototype would make sense primarily on long pages with a lot of text and images. Exactly how the interviewees would use the tool in the future also depended heavily on their profession and education, because this is how they come into contact with different digital products and have different needs.

If this research project would have been repeated, it had been more useful to conduct a pre-test and a post-test with the experts. During the pre-test, the prototype could be refined at a very early stage of development, thus the development could be more focused on the needs of blind users. Once the implemented features are ready for use, the post-test would be used for evaluation, so the existing features can be investigated in practice and fine-tuned if necessary.

6. Future Work

The prototype, that was developed in the course of this master's thesis, served as a proof of concept. Hence, it only provided the necessary functions to allow the blind participants to test it in practice. In this sense, the prototype was specifically designed for Wikipedia articles, because the empirical research showed, that their DOM-trees were the most suitable ones.

In the future, this tools could be further developed to work on many other websites as well and process other types of image-to-text layouts (e.g. two grouped images and one paragraph).

In the prototype the raycasting method recognized images and texts on webpages by matching both elements with a list of criteria. However, on other websites than Wikipedia there are most probably images and texts that do not match the created criteria list, so they will mistakenly not be taken. As a solution, the raycasting method could be combined with AI, in order to make the image and text selection more generic via Large Language Models.

Additionally, the tool could also be released for other devices such as smartphones and tablets. The advantage would be that blind users, who prefer using the internet on their smartphones or tablets, could also benefit from the tool. As a result, the accessibility would be extended and thus the web inclusion would be improved.

Furthermore, the next prototype could have several configuration options that would allow the user to fine-tune the image descriptions more to their personal interests. In addition to that, the function that embeds the image descriptions into the main article's text could be optionally activated.

In the future, the prototype could not only operate as a stand-alone software, but also be rolled out as an add-on for NVDA (or other screen readers). NVDA is an open-source screen reader that can also be customized. The advantage of this would be that only one system would access the websites and there would be no interferences.

Finally, it should be mentioned that blind people must be involved in all further development steps. Therefore, it would make sense not only to involve them through interviews, but also to make this tool accessible to the blind developers community and to develop it together with them.

7. Conclusion

The goal of this master's thesis was to develop a prototype that would allow blind people to have images on websites described in context by using AI. It was important to find out which of the latest versions of AI was best suited to solve this problem. Factors to consider included performance, cost, privacy, and the quality of the generated responses. As a result, the developed prototype was evaluated in practice by conducting a guided interview with five blind experts.

In the course of the literature research, the focus of the work was concentrated on three topics. Firstly, it was explained what the term blindness means and why the inclusion of blind people is important in all areas, but particularly in the web sector. Secondly, it was discussed to what extent the experiences of blind and sighted people differ when navigating websites and perceiving images. And thirdly, the topic of AI was discussed in more detail with a focus on Generative AI, which served as a basis for developing the product.

The way how blind and sighted people use websites varies in some aspects. In particular, blind people use screen readers for navigation and for accessing images. Screen readers are controlled using the keyboard and the selected content from the webpages can either be read aloud via audio output or is displayed on a braille display. In comparison to that, sighted people use their vision for web navigation and viewing images.

Based on the navigation patterns of blind and sighted people, it can be seen that both groups strive for efficiency when navigating websites and therefore simply glide over a lot of content. However, blind people repeatedly experience barriers in this regard, as website developers often forget or do not know how to make elements accessible to screen readers. This often also applies to images on websites, since these can only be perceived by blind people with the help of alternative texts, but the formulated descriptions are often inadequate or do not exist at all. Blind people often develop (personal) navigation strategies to overcome many barriers, but there are also barriers that cannot be conquered by screen readers, so in such cases the assistance of sighted people is needed.

The present work is therefore primarily devoted how Large Language Models can help to make images on websites more accessible for blind people without human assistance. As the research showed, Generative AI in its current form is not a fundamentally new invention. The concepts have existed for many years, but back then the algorithms were not as sophisticated as they are today and both the computing power and the volume of data were not available. It has taken the advancement of just mentioned factors to achieve meaningful results with the help of these technologies. The current Large Language Models, which were discussed in more detail in the course of the research, are the GPT-series from OpenAI, but also LLaMA and LLaVA from Meta. Both model families were theoretically examined for advantages and disadvantages, such as the number of parameters, the maximum context size and the cost. A clear difference between the two companies became apparent: OpenAI's Large Language Models can only be used via ChatGPT or a special software interface, and in some cases for a fee. Meta's models are free and work offline, but users have to cover the cost of the necessary hardware themselves. In the thesis strategies were also proposed that would enable users to optimize the text of their prompts to generate responses of the best possible quality. These findings served as the basis for the prototype to generate qualitative image descriptions.

After an in-depth theoretical research, a prototype was developed in an empirical and iterative setting. In contrast to conventional tools, the underlying concept was to describe images on websites and, in particular, to focus the image descriptions on the relevant image features in respect to the context of the webpage. In addition, the alternative attributes of the selected images should not only be overwritten with the new image descriptions, but the descriptions should be embedded in a contextually relevant place in the main text, in order to avoid a long navigation path with the screen reader.

The development of the prototype was divided into two components. The first component had the task of extracting the text passages and corresponding images from a webpage. Its function was to wrap the extracted data meaningfully in a prompt to allow the AI to generate a precise and contextually relevant image description. After receiving the AI response, the response had to be returned to the first component, that inserted the new descriptions back into the DOM. After a few data extraction attempts with the prototype on various websites, the decision was made to limit the project to Wikipedia articles, to reduce complexity of the project. Wikipedia articles were ideal for the development because they provided the optimal conditions in terms of text passages, images and DOM tree.

In the first empirical research run, an attempt was made to execute this algorithm (extracting, analyzing and modifying) completely with GPT-4o or LLaMA/LLaVA. The first attempt was to extract the necessary data of a webpage via the HTML file. It was expected that the AI's contextual awareness would be beneficial for this task. The first signs of the response looked promising, but then it failed due to hallucinations, which were probably caused by the excessive length of the HTML file with the prompt. On top of that, the estimated costs were so high that a batch process would have led to enormous costs. Subsequently, another experiment was conducted in which the Generative AI had to extract the correct data from a screenshot of the website. The results of this run were better, but it again failed at a certain response length due to hallucinations in the response. Another problem with this attempt was that the screenshot did not provide any connection points to the website so that the generated image descriptions could not be embedded in the correct place in the DOM.

The second and successful attempt to accomplish data extraction using the prototype was based on a method that seemed unusual at the first glance. This method was raycasting, which is usually used in game development. Raycasting was implemented using JavaScript and was intended to simulate the AI's contextual understanding for web pages. It was assumed that text passages and images with similar contexts could be found next to each other within the web layout. Another advantage of implementing raycasting was that this algorithm could be implemented using traditional programming, which completely prevented hallucinations when extracting data. Finally, this method did not incur any costs because no AI was used for it.

For the development of the second component, which was intended to generate the image descriptions, various communication modes and services from OpenAI were tested. At this point, the models from Meta were no longer used because the appropriate infrastructure could not be established for them. The research with the OpenAI services showed that the Assistants API was best suited for the scenario of this work due to its quality and low costs. Having the option to equip the Assistants API with files in advance helped to eliminate redundant content from the prompts and thus keep costs low.

The evaluation of the developed prototype was carried out by guideline-based expert interviews. These experts were five blind people of different age, gender and professional background. Results showed that the interviewees found the prototype helpful. They mentioned that they had not previously been aware of any tool that automatically described images on websites and referred to contextually relevant details. In this sense, the apps BeMyEyes or BeMyAI were often mentioned in a positive way, as they allow blind people to take photographs of their environment first. Both apps will then provide blind users with an image description, however BeMyEyes analyses pictures via humans and BeMyAI uses an AI interface. Both apps were developed with the focus on the real environment, so they do not provide solutions for context-based descriptions for images on websites.

The interviewees confirmed that they would continue to use the prototype, although some of them would prefer to use it in a different way. They expressed wishes for how the prototype could be further developed. For example, there could be both configuration options to adapt the image descriptions better to the personal interests of the users and an interactive chat box where further questions for the AI about an image could be asked. Further points of criticism were that the performance of the AI was too slow and that the data processing at OpenAI would not be transparent, which makes the guarantee of data protection more difficult. At this point, it must be emphasized that the AI does not come close to human intelligence. In the study, sometimes very good image descriptions were created, but others were not well written from the perspective of sighted developers. It must be said that AI reaches its limits when generating image descriptions and does not yet come close to the capabilities of a human being.

The interviewees also made it clear that overcoming barriers is not solely dependent on assistive technologies. They therefore insisted that website developers in particular should improve their awareness and skills with regard to accessible websites. These statements also show how important the research of this thesis is for blind people, even though they only make up a small percentage of the population. Accessibility on websites and all associated multimedia elements not only contribute to their independence and autonomy, but also improves their overall quality of life and equal opportunities. Furthermore, research projects in this field can enhance the democratization and inclusion of blind people in society, by working together with blind and sighted people on solutions. Therefore, this work should end with a statement from an interviewee:

> *"In the field of blindness, you have to distinguish between political practice and lived practice. Political practice is: blind people can't do anything [...] have no education and so on and so forth. Unfortunately, this is still the political message being sold. The lived practice, however, is that more and more of us are graduating from schools and colleges outside of the traditional institutions and concepts. And [...] they are out and free in the world. And that's why it's important to develop such tools with the people who will use them, by involving them in the development, in the projects and in the responsibility. [...] And also that they make knowledge available. [...] I don't think it will be a tool. Rather, you will always need a certain set of tools and, of course, the appropriate training to be able to use such things."* (I 5, 00:10:09-7)

Appendix

Appendix A.

Interview by Mag. phil. Barbara Levc

The interview was originally conducted and recorded in German. For the purposes of this master's thesis, it was translated into English. In this interview, *I* stands for the interviewer and *B* for the interviewee.

I: Could you please describe in a few sentences what kind of visual impairment you have.

B: I have a progressive retinal disease, the medical term is retinitis pigmentosa, which means that the retina, so to speak, gradually, ah, this is the so-called tunnel vision, so the cells of the retina gradually stop working. That means you start with the visual impairment of retinitis and it leads to blindness. That is, from my visual history, it is likely that I saw clearly in my early years and then it became less and less. So I have memories of the visual impressions of things that I still saw and also visual images, but these do not necessarily correspond to reality.

I: Which tools do you use to navigate on websites? (e.g. a screen reader or a braille keyboard)

B: So basically a braille display and screen reader. For the screen reader, now, practically only NVDA and yes, and so, and just the braille display as hardware and the speech output. I actually work in parallel.

I: Do you have another computer at home, or a MacBook or something?

B: No, at the moment I have everything about so-called Windows system, so I have a laptop, a stand PC, so here in the office. At home a laptop, each with the Braille display. I use Apple technology on my cell phone. I have an iPhone and I use the VoiceOver, the speech output. So I swipe to the elements and double-tap to activate elements or also read with gestures, for example a newspaper article, so there are gestures that you can use to have it read the whole thing and you don't have to swipe individually for each destination. Also to switch between menus, for example when you are writing a WhatsApp message and want to correct it, so that you can go back or forward word by word or letter by letter. For example, I can go back word by word. I can also adjust the speed and volume and use all apps, so a lot of them are standard, and then there are also some specific apps for blind and visually impaired people, for example "Seeing AI" or "Be My Eyes" or "Be my AI". These are apps where you can photograph things and get a scene or image described. You can also have short texts read aloud, for example letters or something like that. You can also request human support with the "Be My Eyes" app. There is a network of people who are inside. You can contact a person, the nearest one who is available, and they are then connected to my cell phone camera and can then, for example, describe something or provide information. Last time, for example, I used it to sort Zotterschokolade. I made a larger one and the chocolate bars are the same size. I put them all on the table and then, with the help of a helper, I sorted the right chocolates into the right Easter baskets. Then there is a special navigation system for visually impaired and blind people, which is a pedestrian navigation system called "Blind Square", where you can have a route described, but it also provides information about the surroundings, for example, which shops are in the area, things like that. There are a wide range of special apps, some of which I also use.

I: I've already noticed that there are many innovations and a lot of variety. It's very exciting. What strategies do you use when you access websites that are new to you?

B: I always start by using keyboard shortcuts to go to the place that the screen reader defines as the "top," so to speak, and then I use the arrow keys to go down line by line to get an idea of how much menu there is before I get to the actual content and so on. And anyway, how accessible is the whole thing in principle? That's the first exploration, so to speak.

I: Do you also use a skip link?

B: It may well be that you jump from link to link. But it may also be that I just go through the headlines once. There is a keyboard command for this, so if you use the letter "H" for "Headline", you jump from headline to headline.

I: I once saw on Wikipedia that you can even skip the navigation.

B: Yes, there are jump labels on some websites. These are hidden links at the beginning of the page where you can jump directly to the main content, for example, and then you can skip the entire navigation or you can go directly to the search. These are direct links on the page where you can go directly to the individual areas. Not every website has this, but where it exists, it is very practical. I should also mention that these anchors are sometimes displayed, but they don't always work. For example, on the university website, they sometimes don't work. They are there, but you can't click on them, especially on the older pages. The university website is currently being redesigned, so on the older pages it sometimes doesn't work. So the university is working on it, but all the pages are being redesigned and in the course of this, it is being checked to see if it works.

I: How do you experience different content on websites? Namely:
a. Texts, e.g. length, difficulty of content, typos

B: Well, there are no restrictions on reading aloud. Typographical errors are read aloud, so I hear them immediately. And I actually also use what many sighted people are now using, you are more likely to notice typos when they are read aloud to you than when you see them, so to speak. There is now a read-aloud function, and I know from a colleague who is a lawyer that she has it read to her when she is preparing a legal document, because the building blocks are often too similar or the content is such that she might otherwise overlook errors just by looking at it. I am a very precise typo finder. The only thing it doesn't really pick up well is upper and lower case spelling mistakes, but otherwise everything. So otherwise nothing escapes me. Even with punctuation mistakes, so the intonation goes with it in the emphasis and then you also notice that something is wrong when the intonation suddenly no longer fits. What can happen, happens less often now, but it still happens when the programming code of a website is in a different language than appears on the website. There may well be a German text on it, but if English is actually stored in the programming, then the speech output adjusts to it and reads the German text with an English pronunciation, which is incomprehensible. Conversely, because I don't always change the language, I now have the ability to understand English with German pronunciation, so if I quickly go to an English website and don't want to change the language, it works fine, but it doesn't work the other way around. So if I go to an English website, so you can change the languages in the screen reader, it has a lot of languages with different voices, they also have different names. And if I consciously read something in English, then I consciously change the language in the screen reader. If the website is incorrectly stored, then changing it doesn't help because the screen reader adapts itself and changing the language doesn't help. So, so to speak, the programming code has to store the language that can also be read at the front. It's getting better, but it used to be relatively common. It happens less often now, but I still come across it sometimes.

b. Links z.B.: Recognizability, label, within/outside of continuous text, in the menu

B: So the links, on the one hand, if they are really defined as links, the screen reader will announce them and they will also be marked on the Braille display, and you can jump from link to link using the tab key. Buttons are different, they are also used as switches, and you can't jump from one to the next using the tab key. While a link can always be activated using a screen reader, buttons may not work. Or that only work in combination with a certain browser, for example, or that don't say something to indicate that they're not activated, but then are activated anyway. So with buttons, there are a few variations that don't work 100% barrier-free. Something in the background, the way they're programmed, must be the cause. And then, of course, it is also important that the labeling of both the links and the buttons tells you what is behind them, so if it's not done well, then the speech output either just says "link" or the programming code behind it, or just "switch" and yes. So, for example, when you order something, I accept the terms and conditions, so that it really says that or "next" or "back", so whatever the thing is supposed to do. If there is only an arrow on it, it won't be read to me. So it doesn't have to be visible, but there has to be an alternative text behind it.

c. Images e.g.: alt text, content (living beings, landscapes, objects, diagrams)

B: It depends on the person who creates the alternative text. I don't think there are any real guidelines yet. Automatic alternative texts often don't work at all, because then an image might contain the text. But with the image descriptions or with the alternative texts that are really actively created, it's often just the decision of the person creating the alternative text, based on what they think is important. So if, for example, there are pictures of some events on the university homepage in the news articles, then sometimes it is listed who is in the picture from left to right and just the people and their names, or a picture that just contains some object or something else is described. So it really varies a lot and depends on the picture. We now also adapt study literature for accessibility, including books, presentations, slides, handouts, digital documents, etc. And if there is a relevant image or graphic, we now have our colleagues first create a description of it. They are usually pretty good and may shorten them further because the AI sometimes describes in great detail and not everything is then really 100% important and so that it doesn't become too much, the colleagues then shorten it a bit more. But it makes the work a lot easier now because you let the AI create a draft first and then you can just make the adjustments or improve it yourself.

I: That is to say, because probably when people create it, then people are a bit shortsighted, so to speak, because they take things for granted that are interesting. Is that what you mean?

B: No. The AI also describes things the way it is fed, that is, when an AI describes a landscape to me, it often describes it like an advertising brochure for a vacation. So that can be nice in a private context, but when it comes to image descriptions, it is particularly important that the AI also interprets to some extent, and that must not be the case with a script, for example, a slide for a course, because the interpretation then comes from the lecturer, so the image can really only be described and adapted, so the AI can also be blinkered, so you still need the person to look over it again and see what other information is provided.

d. Videos e.g.: transcription, audio

B: At the moment, there is something called audio description for videos, where a person describes what is happening in the phases where there is no speech in the video itself, where descriptions are recorded. There is an example on the university's diversity website, so there is a website, I think it's called, that's called diversität.uni-graz.at and there are videos on the topic of diversity that were shot in 2016 or 2017. And on the one hand, there are the bloopers that happen to you in everyday life and then there are a few others and there are really three versions. There is one version with nothing, just the video, one with subtitles and one with audio description. This is a very good example where you can compare it well. Otherwise it is still rather rare for videos, but for films there are some, for example TV films, TV series, sports broadcasts with live audio transcription for football matches. So audio description of a football match is amazing, because the regular sports commentary, which only focuses on who has the ball, lists the names and then writes something like "goal" and you don't know how it happened. The audio description really describes what is happening. The people are fantastic because they have to decide incredibly quickly what is important on the field right now and describe that as well. "Passes the ball," "gets a leg up on," or, so that's sort of, not just the results like "free kick," but also how it happens. They have to decide incredibly quickly what to describe and how to describe it. So it's really worth listening to. It creates a sense of suspense. So it's really people doing it, and of course they're enthusiastic about it. They're fully involved. They're football experts and they're fully immersed in the story.

I: Wie ist dann zum Beispiel für Sie Youtube?

B: What is YouTube like for you, for example?

I: Did you tell me that a blind student built a barrier-free Moodle at the TU? What is a big mystery to me is how he coded it blindly.

B: So he had really just reorganized Moodle in a certain way. I don't know how he did it. I was just a test subject to try out how to navigate inside, and that was great because to get to the course content, I saved myself the trouble of going through the three menus at the front. Because Moodle itself is readable, it is just very laboriously organized when you read it with the screen reader, because many menus and buttons come up and then you actually get to the course content. Moodle can do an incredible amount, which is great, but when enthusiastic people want to use everything, it gets exhausting.

I: I would be interested to know what information you would like to obtain from pictures, and what information is useless.

B: Ah, that depends very much on the picture. If it's a landscape, then I'm happy to receive a nice description of the landscape. If it's a, let's say, if it's a website where you buy something, so if it's a shopping website, then I would like to have, for example, if you start from a piece of clothing, a detailed description of the piece of clothing and the pattern and the color scheme, so that I can decide whether I want it or not. If it's a picture of the honorary doctorate ceremony from somewhere at the university, I'm not interested in what color the rector's tie is. I am interested in what can be seen in the picture or what scene is being presented there. If it is a graphic, for example, for statistical data, then I am actually more interested in the result, because if I now have a pie chart, for example, and this pie chart shows that 1% of all students have a disability in the narrower sense or 12% have chronic or mental illnesses and a further 8% have allergies or something something like that, then this information is enough for me, I don't need more information, this area is hatched in the pie area, the next is dotted, the third is somehow, unless there is a task that now tells me: "One third of the circle is black, one third is or one quarter is hatched and the rest is dotted. What percentage is the rest?" If it's a math problem, then I need to know. But if it's just to show results, I don't care about the pie chart, I just need the percentage. So it's really very closely linked to the content.

I: So you can use the AI benefits for such things, as far as they can and fully engage with the context?

B: Exactly, so that's, well, sometimes, when something is only shown in the diagram without any additional explanations, then the additional explanations are important.

I: I found a study, I think it was by Meta, in which they surveyed a whole range of people, from e-commerce to social media. They asked: "What are you interested in when it comes to people? So also irregularities, e.g. a birthmark.

B: I'm interested in the hairstyle and how the people look, so as I said, it depends on the content. If it's some celebrity, then I'm interested in what the person is wearing. But as I said, if it's the principal, then I don't care, because then I can already guess what it is.

I: Is DiCaprio wearing a beard or not more relevant?

B: Yes maybe, but not necessarily now, or has big ears.

I: Da wollten Menschen bei der Studie ableiten, welche Leidenschaften Menschen haben, ob sie vielleicht bei einem Hobby fotografiert wurden, so in diese Richtung.

B: The study was trying to deduce what passions people have, whether they were perhaps photographed during a hobby, something along those lines.

I: It's also a matter of personal interest, what the person wants to know, isn't it?

B: Exactly. There is probably no ideal option because what interests me may not interest someone else and vice versa, so there is probably no "fits-all" option that satisfies everyone.

I: Okay, that means now, with the project I'm doing. For example, I have Der Standard and I have a picture and next to it the text that describes the picture, be it about politics. Then the AI should take from the context from the body text whether the clothing is relevant or not and then go into it in the description of what the picture shows.

B: Yes exactly, with politicians it doesn't really matter, on the contrary, especially with female politicians, I don't want to know, because the AI might be "gendered" because with female politicians, their clothing and appearance are very often commented on and not with male politicians. That is a gender bias and it may well happen that the AI adopts it because the AI is fed by people. But if it says that a particular politician is gesticulating violently during a speech in parliament, then that may be important to know that the picture shows the politician in the middle of a heated debate or something like that.

I: Do you currently have to bear the costs for assistive technologies? If so, how high are they? Would you like more technologies to be provided free of charge or for the costs to be borne by someone else?

B: That depends very much on the system and my position in life. Well, for one thing, NVDA is free software, which is great. You're always invited to donate because they're also developing it further, and that's really important, not only for me personally, but also deliberately kept that way for people in countries where there are no social benefits or social benefits that wouldn't pay for something like that. Apple has simply integrated these things by default, so they are included for free, and Apple products are generally more expensive. Regarding the rest, there is the screen reader software Jaws, which is relatively new and can be better adapted for special systems, so if someone has to work with a very specific program in a company, then SAP, for example, can be customized for Jaws. Otherwise, NVDA can do everything in standard operation, and some things even better, in my opinion. And whatever costs a lot is the hardware, so the braille display, the magnifying screen, and there it is, if you are employed and need it for your job, then the costs are covered by public funding. If you need it for private purposes, if you are retired or if you are in vocational training, it will also be covered. However, if you are employed and want to further educate yourself, it is a gray area and often only partially covered, e.g. 30%.

I: It was a spur-of-the-moment idea. What if the government were obliged to sponsor each person? Would that make sense?

B: I think it would, so the whole thing has to do with the fact that our funding system is so fragmented between the federal and state governments. For the vocational sector, it is the federal laws and for the non-vocational sector, it is the state laws, which also vary from state to state. And the ideal solution would be if the entire area of care with assistive technologies, etc. would simply be taken away from the federal states to the federal government, because then it would be standardized and then, for example, any taxes I'm not exactly familiar with tax law, but I assume that some taxes are levied for some digital things, so that at least part of it is earmarked for assistive technologies for people with different disabilities. This doesn't have to be just blindness; there are also eye or mouth controls for people with severe physical disabilities. It would make sense to earmark some of the money for this and to ensure that all such technologies are free of charge for everyone.

I: Thank you very much for this interview!

B: With great pleasure!

Appendix B.

Interview

B.1. Interview Guideline

Demographics

1. How old are you and what do you do for a living?
2. Could you please describe your visual impairment in a few sentences?

Websites

1. Which devices and tools do you use to navigate websites?
2. What are the main problems you encounter when navigating websites?
3. What strategies do you use to overcome these problems?
4. What has been your experience so far with accessing visual content on websites? How helpful are alt texts? What information is useful and what is useless?

Prototype Demonstration

Evaluation

1. Would you continue to use the tool? If yes, why and for what? If not, why not?
2. What else would you like to see in this tool?
3. What other tools would you need to make navigation easier and make images more accessible??
4. Do you think this prototype could be significant for the future, with sophisticated technologies that act faster and deliver more accurate content? If so, why? If not, why not?

Future

1. What tools do you think might be developed in the future to make images more accessible and improve overall navigation?

B.2. Declaration of Consent

The privacy policy is attached in the original, which was written in German:

Information zum Datenschutz und zur Einwilligungserklärung zur Erhebung und Verarbeitung personenbezogener Daten im Kontext von Masterarbeiten

Datenschutzerklärung

Die TU Graz nimmt den Datenschutz sehr ernst und behandelt personenbezogene Daten vertraulich und entsprechend den gesetzlichen Vorschriften. Diese Datenschutzerklärung soll gem. Art 12, 13 DS-GVO über Zweck, Rechtsgrundlage und über die Rechte iZm der Verarbeitung personenbezogener Daten durch die TU Graz, Inffeldgasse 25D, 8010 Graz informieren.

Verwendung der personenbezogenen Daten der befragten Person

Die personenbezogenen Daten werden für die Masterarbeit erhoben, verarbeitet und gespeichert. Die Aufzeichnung erfolgt durch eine Audio- und/oder Videoaufnahme. Zur Auswertung wird ggf. ein schriftliches Protokoll (Transkript) angefertigt. Im Transkript werden die Angaben zur Person anonymisiert.

Speicherdauer

Die personenbezogenen Daten bleiben bis zum Widerruf der Einwilligung gespeichert. Darüber hinaus werden nur die unbedingt notwendigen personenbezogenen Daten zum Zwecke des Nachweises der Einwilligung bzw. des Widerrufs für die Dauer von max. 3 Jahren gespeichert. Die Daten werden nicht an externe Empfänger:innen übermittelt.

Die Rechte der beforschten Personen

Im Zusammenhang mit der Verarbeitung von personenbezogenen Daten verfügen die beforschten Personen jederzeit über die folgenden Rechte, die bei den durchführenden Personen geltend gemacht werden können:

- Recht auf Auskunft über die betreffenden personenbezogenen Daten (Art 15 DSGVO), Recht auf Berichtigung (Art 16 DS-GVO) oder Löschung (Art 17 DS-GVO) oder auf Einschränkung der Verarbeitung (Art 18 DS-GVO),
- Recht auf Datenübertragbarkeit (Art 20 DS-GVO),
- Recht auf Widerspruch (Art 21 DS-GVO),
- Recht auf Widerruf der Einwilligung (Art 7 Abs 3 DS-GVO), wodurch die Rechtmäßigkeit der bis zum Widerruf erfolgten Verarbeitung nicht berührt wird.

Darüber hinaus besteht das Recht auf Beschwerde (Art 77 DS-GVO), welches bei der österreichischen Datenschutzbehörde, Wickenburggasse 8, 1080 Wien, Telefon: +43 1 52 152-0, E-Mail: dsb@dsb.gv.at als zuständige Aufsichtsbehörde einzubringen wäre.

Vertraulichkeit und Anonymisierung der personenbezogenen Daten

Die Vertraulichkeit all der im Rahmen der Masterarbeit erhobenen personenbezogenen Daten ist gewährleistet. Persönliche Informationen werden nicht an dritte Personen weitergegeben. Es können Ergebnisse im Rahmen von Vorträgen öffentlich vorgestellt oder im Rahmen von Lehrveranstaltungen verwendet werden. Bei jedweder öffentlichen Verwendung bzw. Ergebnisdarstellung werden die Informationen ausschließlich in anonymisierter Form veröffentlicht.

Ich habe das Informationsblatt gelesen und verstanden.

Einwilligungserklärung zur Erhebung und Verarbeitung personenbezogener Daten im Kontext von Masterarbeiten

Hiermit erkläre ich mein Einverständnis zur wissenschaftlichen Verwendung meiner im Rahmen der Masterarbeit von David Mischak, MA generierten Daten.

Für die wissenschaftliche Auswertung der von mir zur Verfügung gestellten Daten werden alle Angaben, die zu einer Identifizierung meiner oder anderer Personen führen könnten, verändert oder entfernt.

Ich kann diese Einwilligung mit Wirkung auf die Zukunft jederzeit schriftlich bei David Mischak, MA widerrufen.

Ich bestätige mit meiner Unterschrift, dass ich freiwillig an diesem Projekt teilnehme über den Zweck der Datenerhebung aufgeklärt worden bin.

Bibliography

Al-Harbi, S. (2019). Language development and acquisition in early childhood. *Journal of Education and Learning (EduLearn)*, *14*, 69. https://doi.org/10.11591/edulearn.v14i1.14209 (cit. on p. 28).

American Foundation for the Blind. (2024, February). *Key Definitions of Statistical Terms*. https://www.afb.org/research-and-initiatives/statistics/key-definitions-statistical-terms (cit. on p. 9).

Apple Inc. (n.d-a). *Verwenden von VoiceOver zum Navigieren auf Webseiten mithilfe des DOM- oder Gruppenmodus auf dem Mac*. Retrieved July 30, 2024, from https://support.apple.com/de-de/guide/voiceover/vo2711/mac (cit. on p. 16).

Apple Inc. (n.d-b). *VoiceOver – Benutzerhandbuch*. Retrieved July 30, 2024, from https://support.apple.com/de-de/guide/voiceover/welcome/mac (cit. on p. 16).

Assistants api overview. (n.d.). OpenAI. San Francisco, CA. Retrieved November 12, 2024, from https://platform.openai.com/docs/assistants/overview (cit. on p. 71).

Awan, A. A. (2024a, May). *How to run llama 3 locally: A complete guide*. Retrieved July 15, 2024, from https://www.datacamp.com/tutorial/run-llama-3-locally (cit. on p. 32).

Awan, A. A. (2024b, May). *Fine-tuning llama 3 and using it locally: A step-by-step guide*. Retrieved October 17, 2024, from https://www.datacamp.com/tutorial/llama3-fine-tuning-locally (cit. on pp. xvi, 33).

Baxter, K., & Schlesinger, Y. (2023, June). *Managing the Risks of Generative AI*. Retrieved October 20, 2024, from https://hbr.org/2023/06/managing-the-risks-of-generative-ai (cit. on p. 36).

Bhowmick, A., & Hazarika, S. M. (2017). An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. *Journal on Multimodal User Interfaces*, *11*(2), 149–172. https://doi.org/https://doi.org/10.1007/s12193-016-0235-6 (cit. on p. 15).

Bonvillain, N. (2020). *Language, culture, and communication: the meaning of messages* (8th edition.). Rowman Littlefield. (Cit. on p. 28).

Bortz, J. (2009). *Forschungsmethoden und evaluation für human- und sozialwissenschaftler : Mit ... 87 tabellen* (4., überarb. Aufl., Nachdr..). (Cit. on p. 89).

BSVÖ. (n.d.). *Statistische Daten*. https://www.blindenverband.at/de/information/augengesundheit/97/Statistische-Daten (cit. on p. 6).

Carolan, K., Fennelly, L., & Smeaton, A. F. (2024). A review of multi-modal large language and vision models. https://arxiv.org/abs/2404.01322 (cit. on p. 31).

Caruana, A., & Goater, K. (2024, October). *The dark side of generative AI: How using AI in marketing, advertising and communications can expose confidential organisational information*. Retrieved October 20, 2024, from https://mumbrella.com.au/the-dark-side-of-generative-ai-how-using-ai-in-marketing-advertising-and-communications-can-expose-confidential-organisational-information-851942 (cit. on p. 36).

Coutinho, C. (2023). Raycasting. In *Roblox lua scripting essentials* (pp. 279–304). Apress L. P. (Cit. on p. 58).

Cruz, A. (2006, September). A blind man is led by his guide dog in brasília, brazil. Retrieved November 11, 2024, from https://commons.wikimedia.org/wiki/File:Caoguia2006.jpg (cit. on pp. xv, 67).

Derner, E., & Batistič, K. (2023). Beyond the safeguards: Exploring the security risks of chatgpt. https://arxiv.org/abs/2305.08005 (cit. on pp. 36, 107).

Dop, T. (2020). Section 2: Fundamentals of natural language processing. In *Hands-on natural language processing with pytorch 1. x*. Packt Publishing, Limited. (Cit. on p. 30).

El Amri, A. (2024). Llm prompt engineering for developers : The art and science of unlocking llms' true potential /. (Cit. on pp. 29, 31, 33, 34).

Firth, A. (2019). Practical web inclusion and accessibility : A comprehensive guide to access needs / (1st ed. 2019..). (Cit. on p. 81).

Firth, A. (2024). *Practical web accessibility: A comprehensive guide to digital inclusion* (Second edition). Apress. (Cit. on pp. xiv, 2, 11–15, 105, 106).

Flaherty, K. (2017, November). *Zigzag Image–Text Layouts Make Scanning Less Efficient*. https://www.nngroup.com/articles/zigzag-page-layout/ (cit. on p. 23).

Forrest, S. L., Mercado, C. L., Engmann, C. M., Stacey, A. W., Hariharan, L., Khan, S., & Cabrera, M. T. (2023). Does the Current Global Health Agenda Lack Vision? *Global Health: Science and Practice*, *11*(1). https://doi.org/10.9745/GHSP-D-22-00091 (cit. on pp. 6, 7).

Foster, D. (2023). *Generative deep learning, 2nd edition*. O'Reilly Media, Inc. (Cit. on pp. xiv, 2, 29, 31, 105).

FreedomScientific. (n.d.). *JAWS*. Retrieved July 30, 2024, from https://www.freedomscientific.com/products/software/jaws/ (cit. on p. 16).

Fumo, J. (2017). Why is everyone talking about artificial intelligence? Retrieved July 19, 2024, from https://towardsdatascience.com/why-is-everyone-talking-about-ai-73bab31bf9c1 (cit. on p. 27).

Gläser, J. (2006). *Experteninterviews und qualitative inhaltsanalyse : Als instrumente rekonstruierender untersuchungen* (2., durchges. Aufl..). (Cit. on pp. 87, 92).

Gläser, J. (2009). *Experteninterviews und qualitative inhaltsanalyse : Als instrumente rekonstruierender untersuchungen* (3., überarb. Aufl..). (Cit. on p. 88).

GNOME. (n.d.-a). *Orca*. Retrieved July 30, 2024, from https://orca.gnome.org/ (cit. on p. 16).
GNOME. (n.d.-b). *Orca Screen Reader*. Retrieved July 30, 2024, from https://gnome.pages.gitlab.gnome.org/orca/help/ (cit. on p. 16).
Google. (2022, October). *Hello world extension*. Google. Mountain View, CA. Retrieved November 16, 2024, from https://developer.chrome.com/docs/extensions/get-started/tutorial/hello-world (cit. on p. 77).
Henry, S. L. (2024, March). *Introduction to web accessibility*. Retrieved May 27, 2024, from https://www.w3.org/WAI/fundamentals/accessibility-intro/#what (cit. on p. 11).
Henry, S. L., Abou-Zahra, S., & White, K. (2016, May). *Accessibility, Usability, and Inclusion*. Retrieved August 27, 2024, from https://www.w3.org/WAI/fundamentals/accessibility-usability-inclusion/ (cit. on pp. 10, 11).
Hustak, T., & Krejcar, O. (2016, January). Principles of usability in human-computer interaction. https://doi.org/10.1007/978-3-662-47895-0_7 (cit. on p. 11).
Jones, R. (2018, January). *A Tale of Two Rooms: Understanding screen reader navigation*. Retrieved June 2, 2024, from https://www.tpgi.com/a-tale-of-two-rooms-understanding-screen-reader-navigation/ (cit. on p. 16).
Kosinski, M., & Forrest, A. (2024, March). *What is a prompt injection attack?* Retrieved October 20, 2024, from https://www.ibm.com/topics/prompt-injection (cit. on pp. 36, 107).
Kulkarni, A., Shivananda, A., Kulkarni, A., & Gudivada, D. (2023). *Applied generative ai for beginners: Practical knowledge on diffusion models, chatgpt, and other llms* (1st ed.). Apress L. P. (Cit. on pp. xiv, 27–30, 106).
Lang, M., Heyl, V., & Grevnig, H. (2020, October). *Pädagogik bei Blindheit und Sehbehinderung*. (Cit. on pp. 6, 7).
Marr, B. (2023, May). *A Short History Of ChatGPT: How We Got To Where We Are Today*. Retrieved December 20, 2024, from https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/ (cit. on p. 2).
Mayring, P. (2015). *Qualitative inhaltsanalyse : Grundlagen und techniken* (12., überarbeitete Auflage.). (Cit. on pp. 93–95).
Mayring, P., & Brunner, E. (2013, June). Qualitative inhaltsanalyse. In B. Friebertshäuser & A. Prengel (Eds.), *Handbuch qualitative forschung in der erziehungswissenschaft* (4th ed.). Juventa. (Cit. on p. 93).
Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation* (Fourth edition.). Wiley. (Cit. on p. 87).
Meta. (2023, February). *Introducing llama: A foundational, 65-billion-parameter large language model*. Retrieved July 16, 2024, from https://ai.meta.com/blog/large-language-model-llama-meta-ai/ (cit. on p. 32).
Meta AI. (2024a). *Llama-3:70b* [Large Language Model]. Retrieved July 10, 2024, from https://ollama.com/library/llama3:70b (cit. on p. 51).

Meta AI. (2024b). *Llava-3:34b* [Large Language Model]. Retrieved July 13, 2024, from https://ollama.com/library/llava:34b (cit. on p. 55).

Michaelsexton2003. (2024a). Marie curie — Wikipedia, the free encyclopedia. Retrieved November 17, 2024, from https://en.wikipedia.org/w/index.php?title=Marie_Curie&oldid=1257756082 (cit. on pp. xv, 82, 83).

Michaelsexton2003. (2024b). Marie curie — Wikipedia, the free encyclopedia. Retrieved November 13, 2024, from https://en.wikipedia.org/w/index.php?title=Marie_Curie&oldid=1257756082 (cit. on p. 52).

Moran, K. (2017, September). *Heatmap Visualizations from Signifier Eyetracking Experiment*. https://www.nngroup.com/articles/heatmap-visualizations-signifiers/ (cit. on p. 20).

National Eye Institute. (2023, November). *Refractive Errors*. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/refractive-errors (cit. on p. 8).

Nature, S. (2023, January). *Tools such as ChatGPT threaten transparent science; here are our ground rules for their use*. https://doi.org/https://doi.org/10.1038/d41586-023-00191-1 (cit. on p. 37).

Niebert, K., & Gropengießer, H. (2014). Leitfadengestützte interviews. In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen forschung* (pp. 121–132). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37827-0_10 (cit. on p. 88).

Nielsen, J. (2012, January). *Usability 101: Introduction to usability*. Retrieved May 30, 2024, from https://www.nngroup.com/articles/usability-101-introduction-to-usability/ (cit. on p. 11).

Nielsen, J., & Pernice, K. (2010). *Eyetracking Web Usability* (1st edition.). New Riders. (Cit. on pp. 18, 19, 23, 24, 105).

NV Access. (n.d.). *Features*. Retrieved July 30, 2024, from https://www.nvaccess.org/about-nvda/ (cit. on p. 15).

NVIDIA. (2019). *Data Sheet: Quadro RTX 6000*. Retrieved November 6, 2024, from https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/quadro-rtx-6000-us-nvidia-704093-r4-web.pdf (cit. on pp. xvi, 47).

O'Connor, S., & Liu, H. (2024). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI & SOCIETY*, *39*(4), 2045–2057. https://doi.org/10.1007/s00146-023-01675-4 (cit. on p. 37).

Ollama. (2024). *Llava*. Retrieved October 17, 2024, from https://ollama.com/library/llava (cit. on pp. xvi, 33).

OpenAI. (2024a). *Chatgpt (aug 06 2024)* [Large Language Model]. Retrieved November 16, 2024, from https://chat.openai.com/chat (cit. on pp. 68, 70, 72).

OpenAI. (2024b). *Chatgpt (may 13 2024)* [Large Language Model]. Retrieved July 10, 2024, from https://chat.openai.com/chat (cit. on p. 50).

OpenAI. (2024c). *Chatgpt (may 13 2024)* [Large Language Model]. Retrieved July 13, 2024, from https://chat.openai.com/chat (cit. on p. 54).
OpenAI. (2024d). *Chatgpt (may 13 2024)* [Large Language Model]. Retrieved November 16, 2024, from https://chat.openai.com/chat (cit. on p. 84).
OpenAI. (n.d.-a). *How can I access GPT-4, GPT-4 Turbo and GPT-4o?* Retrieved July 15, 2024, from https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4-gpt-4-turbo-and-gpt-4o (cit. on p. 31).
OpenAI. (n.d.-b). *Models*. Retrieved July 15, 2024, from https://platform.openai.com/docs/models (cit. on p. 31).
OpenAI. (n.d.-c). *Prompt engineering*. Retrieved July 16, 2024, from https://platform.openai.com/docs/guides/prompt-engineering (cit. on p. 34).
OpenJS Foundation. (n.d.-a). *Express: Fast, unopinionated, minimalist web framework for node.js*. OpenJS Foundation. San Francisco, CA. Retrieved November 16, 2024, from https://expressjs.com/ (cit. on p. 78).
OpenJS Foundation. (n.d.-b). *Run javascript everywhere*. OpenJS Foundation. San Francisco, CA. Retrieved November 16, 2024, from https://nodejs.org/en (cit. on p. 79).
Österreichische Arbeitsgemeinschaft für Rehabilitation. (2010). *Bericht zur Umsetzung der UN-Konvention über die Rechte von Menschen mit Behinderungen in Österreich anlässlich des 1. Staatenberichtsverfahrens vor dem UN-Ausschuss über die Rechte von Menschen mit Behinderungen*. https://www.behindertenrat.at/wp-content/uploads/2017/12/2013-BR-ZivilgesBericht.pdf (cit. on p. 10).
Österreichischer Behindertenrat. (2018). *Bericht zur Umsetzung der UN-Konvention über die Rechte von Menschen mit Behinderungen in Österreich anlässlich des 2. Staatenberichtsverfahrens vor dem UN-Ausschuss über die Rechte von Menschen mit Behinderungen*. https://www.behindertenrat.at/wp-content/uploads/2018/07/2018-07-17-ZGB-Deutsch.pdf (cit. on p. 9).
Oyewole, B. (2019, February). *How visually impaired people navigate the web*. Retrieved April 8, 2024, from https://uxdesign.cc/how-visually-impaired-people-navigate-the-web-7f9eab9d9c37 (cit. on pp. 2, 17, 20).
Ozdemir, S. (2023). *Quick start guide to large language models: Strategies and best practices for using chatgpt and other llms*. Addison-Wesley Professional. (Cit. on p. 30).
Paradkar, S. (2017). Mastering non-functional requirements : Analysis, architecture, and assessment / (1st edition.). (Cit. on p. 45).
Pernice, K. (2019, August). *Text scanning patterns: Eyetracking evidence*. Retrieved June 2, 2024, from https://www.nngroup.com/articles/text-scanning-patterns-eyetracking/ (cit. on p. 20).
Persson, H., Åhman, H., Yngling, A. A., & Gulliksen, J. (2015). Universal design, inclusive design, accessible design, design for all: Different concepts—one goal? on the concept of accessibility—historical, methodological and

philosophical aspects. *Universal Access in the Information Society*, *14*(4), 505–526. https://doi.org/10.1007/s10209-014-0358-z (cit. on p. 10).

Porter, D. (2022, January). *Visual Acuity*. https://www.aao.org/EYE-HEALTH/TIPS-PREVENTION/visual-acuity-3 (cit. on p. 8).

Ray, P. L., Cox, A. P., Jensen, M., Allen, T., Duncan, W., & Diehl, A. D. (2016). Representing vision and blindness. *Journal of Biomedical Semantics*, *7*(1), 15. https://doi.org/10.1186/s13326-016-0058-0 (cit. on pp. 7, 8).

Sarkar, D. (2019). Natural language processing basics. In *Text analytics with python* (pp. 1–68). Apress L. P. (Cit. on p. 28).

Schaffer, H., & Schaffer, F. (2019). *Empirische methoden für soziale berufe: Eine anwendungsorientierte einführung in die qualitative und quantitative sozialforschung* (1. Auflage). Lambertus-Verlag. (Cit. on p. 87).

Schwedux, T. (2023). What is chatgpt and why is it so popular? Retrieved July 15, 2024, from https://medium.com/@t.schwedux/what-is-chatgpt-and-why-is-it-so-popular-f1e053225811 (cit. on p. 31).

Shadow. (n.d.). *Regeln und Einschränkungen für Shadow*. https://support.shadow.tech/de/articles/minimum-system-requirements-and-faqs/rules-and-restrictions-on-shadow/64e3100f6593fd01eb5e8288 (cit. on p. 65).

Shadow. (2024). Shadow pc | pro. https://shadow.tech/en-DE/shadowpro/offers (cit. on pp. xvi, 46, 47).

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: Definition and background. In *Mission ai: The new system technology* (pp. 15–41). Springer International Publishing. https://doi.org/10.1007/978-3-031-21448-6_2 (cit. on p. 27).

Siebert, J. (2024, September). *Halluzinationen von generativer KI und großen Sprachmodellen (LLMs)*. Retrieved October 20, 2024, from https://www.iese.fraunhofer.de/blog/halluzinationen-generative-ki-llm/ (cit. on pp. 36, 37).

Stangl, A., Morris, M. R., & Gurari, D. (2020, April). *"person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions*. https://www.microsoft.com/en-us/research/publication/person-shoes-tree-is-the-person-naked-what-people-with-vision-impairments-want-in-image-descriptions/ (cit. on pp. 2, 24–26, 106).

Stigler, H., & Reicher, H. (2012). Praxisbuch empirische sozialforschung in den erziehungs- und bildungswissenschaften (2., aktualisierte und erweiterte Auflage.). (Cit. on pp. 88, 89).

Sulong, S., & Yusof, R. J. R. (2024). Developing a blind user mental model (BlUMM) for web browsing. *Universal Access in the Information Society*, *23*(3), 1343–1367. https://doi.org/https://doi.org/10.1007/s10209-023-01035-5 (cit. on pp. 2, 20, 21).

Susheela, H., Yashwant, S., Vandana Mohindru, S., Manu, S., & Sandeep, D. (2023). Functional and non-functional requirements in agile software development. In *Agile software development*. John Wiley Sons, Incorporated. (Cit. on pp. 43, 44).

Tiwari, A. (2023, November). *Can screen readers read images?* Retrieved June 2, 2024, from https://allyant.com/can-screen-readers-read-images-text-on-images/ (cit. on p. 24).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288 (cit. on pp. xvi, 33).

Turbert, D. (2022, March). *Visual Field Test*. https://www.aao.org/eye-health/tips-prevention/visual-field-testing (cit. on p. 8).

United Nations. (2012, December). *Convention on the Rights of Persons with Disabilities*. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities (cit. on p. 9).

User-duck. (2024). Visual impairment — Wikipedia, the free encyclopedia. Retrieved November 13, 2024, from https://en.wikipedia.org/w/index.php?title=Visual_impairment&oldid=1251102775 (cit. on pp. xiv, xv, 42, 43, 60–62, 65).

Walthes, R. (2022, January). 3 Blindheit – Sehbeeinträchtigung. In *Einführung in die Pädagogik bei Blindheit und Sehbeeinträchtigung* (pp. 52–105). https://elibrary.utb.de/doi/abs/10.36198/9783838558325-52-105 (cit. on pp. 7, 8).

Witt, A. (2024, February). *Computer hardware required to run llama ai model locally (gpu, cpu, ram, ssd)*. Retrieved July 15, 2024, from https://www.hardware-corner.net/guides/computer-to-run-llama-ai-model/ (cit. on p. 32).

World Health Organisation. (2024). *International Classification of Functioning, Disability and Health (ICF)*. https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health (cit. on p. 8).

World Health Organization. (2023, August). *Blindness and vision impairment*. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (cit. on pp. 6, 7).