A novel approach to automating context-driven alternative text generation through purposeful games

A Thesis Submitted for the
Degree of Doctor of Philosophy

By

Nikolaos Droutsas

Computer Science, Brunel University London

2026

Abstract

Accessibility of the Web is a pervasive issue, owing to the persistence of accessibility barriers (e.g., poor navigation, lack of/unsuitable alternative text (alt text), complex web forms), with significant impact on users with disabilities. Alt text barriers in particular, are some of the most prevalent web accessibility barriers affecting a wide range of media and are underpinned by a lack of understanding and guidelines on what constitutes suitable alt text. Past work has shown that the 'context in which the image is used in' is ironclad for suitability yet loosely defined. Whilst there is a need to automate alt text generation, current solutions disregard context in alt text and are lacklustre with regard to suitability. In this research, an empirical exploratory study that investigates the views of web content creators and visually impaired users on suitability is conducted to bridge the functional gap between experiences and best available practice. The first definition of 'Alt Text Context' is proposed providing a systematic way to assess when alt text is necessary and what it should convey. Further, the first crowdsourcing game for context-driven alt text authorship and evaluation—TagALTlong—is presented. TagALTlong's design is informed by relevant literature, empirical qualitative insights and the proposed definition of context in alt text. Following an empirical user study, 125 non-expert players were recruited to play TagALTlong over a six-week period, resulting in 1208 authored and 1836 rated alt text descriptions, respectively. The resulting dataset was used to fine-tune and train an AI model for automated alt text generation to assess whether average human-level alt text quality can be approximated whilst automating the process. Results indicated the improved performance of the model that was fine-tuned and trained on the GWAP-generated dataset compared to pure image processing, subsequently demonstrating the value of the dataset.

Acknowledgements

First, I would like to direct a big thanks to *Dr. Fotios Spyridonis*, my principal supervisor and friend without whom, this journey and work would not have been the same—Fotis was there every step of the way—it was not only once that he worked for the excellence of this work a lot more than I did, constantly steering my keenness to deliver to the fullest. He also believed in me to take on further duties within academia, including teaching and research projects, from very early on in my PhD journey—I owe a lot to Fotis that a small section like this can certainly not cover; yet, the biggest thank you expressed herein is most rightfully earned by him!

I would also like to thank my second supervisor *Dr. Damon Daylamani-Zad*, and *Philipp* Bibik, in collaboration with whom the concept, approach and experiment for the first and second AI models were co-created, respectively, and for supporting the development and deployment.

I would also like to thank my research development advisor *Prof. George Ghinea* for his kind patience and support, and for always being there when I needed him.

Next, I would like to thank my referees, *Prof. Doris Rusch* and *Hon. Prof. Richard Bartle*, who helped me in my journey to pursue doctoral studies. I very much doubt I would have been as resolute in my search for PhD programs if not for Richard's encouraging words and continuous support throughout this journey, nor without knowing that Doris has been rooting for me and was certain that I would make something great out of this PhD journey.

I would also like to direct my sincere thanks to my colleague and friend *Dr. Zear Ibrahim* for his advice in setting up the Cloud infrastructure of the solution proposed in this thesis.

A big thanks is also due to *Prof. Kate Hone* without whom I wouldn't have been able to find thee—most notorious—'player 19'! Thank you Kate!!

I would also like to thank all the *people participating in this research*, and the *institutions and organisations that helped me recruit participants*, including the Royal National Institute of Blind People (UK), WebAIM, AbilityNet, the National Federation of the Blind (Greece), Silktide, KreativeInc Agency Ltd, and Scope.

Mom and Dad – I guess I've found another scholarly chance to thank you for—everything.

Author's Declaration

I, Nikolaos Droutsas, confirm that this thesis and the research conducted therein is my own original work. Research within this work that has been carried out in collaboration with others, as well as material in this thesis that are contained in published papers are acknowledged below, with my contributions indicated.

Publications

- Droutsas, N., Spyridonis, F., Daylamani-Zad, D. and Ghinea, G. (2025b) 'Web Accessibility Barriers and their Cross-disability Impact in eSystems: A Scoping Review', *Computer Standards & Interfaces*, 92, p. 103923. Available at: https://doi.org/10.1016/j.csi.2024.103923.
  Chapters 2 and 3 of this thesis relate to material contained in this publication, including related contributions, limitations and suggested avenues for future work.

- Droutsas, N., Spyridonis, F., Daylamani-Zad, D. and Ghinea, G. (2025a) 'Flower in the Mirror, Moon on the Water: Bridging Perspectives on Alternative Text and Recommendations for Practice', *International Journal of Human–Computer Interaction*, pp. 1–21. Available at: https://doi.org/10.1080/10447318.2025.2499659.
  Chapters 4 and 8, and Appendices A and B of this thesis relate to material contained in this publication, including related contributions, limitations and suggested avenues for future work.

Status of unpublished papers

- Paper title: Painting dragons and dotting the eyes: A context-driven game-based approach to alt text annotation and evaluation — Submitted to: International Journal of Human-Computer Studies (Status: 'Under Review')
  Chapters 5, 9 and 10 of this thesis relate to material contained in this paper, including related contributions, limitations and suggested avenues for future work.

- Paper title: Improving Automated Alt Text Generation with a Human-curated Context-driven Dataset — Submitted to: Behaviour & Information Technology (Status: 'Submitted')
  Chapter 11 of this thesis relates to material contained in this paper, including related contributions, limitations and suggested avenues for future work.

Table of Contents

List of Figures

Figure **Page**

List of Tables

Table **Page**

List of Abbreviations (alphabetically ordered)

Abbreviation	Full term
AI	Artificial Intelligence
ALP	Average Lifetime Play
Alt text	Alternative Text
altC	Alt Text Context
AW	Affect Weight
BBC	British Broadcasting Corporation
CIDEr	Consensus-based Image Description Evaluation
CpI	Cost per Item
CpJ	Cost per Judgement
CxDAI	Cross-Disability Accessibility Impact
DPO	Direct Preference Optimization
DR	Disturbance Rate
GWAP	Game-With-A-Purpose
HCI	Human–Computer Interaction
HomP	Website Homepages
IRR	Inter-Rater Reliability
IntP	Interior Pages
LTJ	Lifetime Judgements
MS COCO	Microsoft Common Objects in COntext
NLP	Natural Language Processing
RLHF	Reinforcement Learning for Human Feedback
RQ	Research Question
SD	Standard Deviation
SoTA	State of the Art
V2L	Vision-to-Language
VIU	Visually Impaired User
VQAv2	Visual Question Answering v2
WA	Web Activity
WAI	Web Accessibility Initiative
WCAG	Web Content Accessibility Guidelines
WCC	Web Content Creator

Glossary (alphabetically ordered)

Term and brief definition

Accessibility Barriers: “any obstacles that prevent individual, especially those with disabilities, from accessing or interacting with online content effectively” (Halpin, 2025, para. 2).

Accessible Design: A design approach that relies on extending standard design principles to people with some type of functioning limitation (ISO and Guide, 2001).

Alternative Text (Alt Text): “a textual substitute for non-text content in Web pages” (WebAIM, 2021), which is accessed via screen readers.

Alt Text Context (altC): A structured semantic definition accounting for multiple factors (Image Type, Webpage Topic, Webpage Purpose, Image Function, Image Intent) that influence how an image should be described in alt text (Section 4.3.1).

Barrier-Free Design: A design approach focusing mainly on lifting barriers for specific individuals in order to perceive an environment as barrier-free (Persson et al., 2015).

ContextALT-O-matic: The second AI model proposed in the thesis, designed to learn to automatically generate more context-driven alt text descriptions based on the GWAP-generated dataset and the use of context prompts (Section 9.2.3.2).

Cross-Disability Accessibility Impact (CxDAI): A proposed measure to quantify the impact of each web content accessibility barrier across different disabilities, calculated using disturbance rates, web activity, and affect weight (Section 3.3).

Crowdsourcing: “the act of taking a task traditionally performed by a designated agent (such as an employee or a contractor) and outsourcing it by making an open call to an undefined but large group of people” (Howe, 2008, p. 1).

Decorative (Eye Candy) Images: An image that is uninformative (i.e., either only used as eye candy or as a duplicate of neighbouring text content) (Silktide, 2020).

Direct Preference Optimization (DPO): A contrastive preference learning technique used to fine-tune AI models by steering them toward human-preferred outputs using pairs of higher- and lower-rated alt texts (Rafailov *et al.*, 2023).

Disturbance Rate (DR): A measure per disability, inspired by past reports on users’ disturbance by accessibility barriers, used in calculating the CxDAI (Berger *et al*., 2010).

Games-With-A-Purpose (GWAPs): Computer games that people play and "could, without consciously doing so, simultaneously solve large-scale problems" (Von Ahn, 2006, p. 92).

HumanALT-O-matic: The first AI model proposed in the thesis, which aims to learn to automatically generate alt text comparable to average human-level quality based on the data generated by players through the GWAP (Section 9.2.3.1).

Human Computation: Short tasks where expertise on a subject matter is not expected and are thus easy for humans but hard to automate for computers (Quinn and Bederson, 2011).

Inclusive Design: An attitude to design that seeks to evolve based on continuous gathering of insights about population diversity, diverse experiences, and interactions within the environment (Persson *et al*., 2015).

Population Diversity: The consideration of a multitude of abilities and contexts (Waller *et al.*, 2015; Heylighen, Van der Linden and Van Steenwinkel, 2017).

Redundancy Mechanism: A GWAP design feature that assigns the same annotation task to many players to ensure consistency and capture diverse opinions (Table 9).

Screen Reader: Assistive software that reads out loud content displayed on computer screens (Dobransky and Hargittai, 2016).

Suitable Alt Text: Alt text that is accurate, complete and concise in relation to the context in which the image it substitutes is used in (Mack *et al.*, 2021).

TagALTlong: The first GWAP for context-driven alt text authorship and evaluation presented in the thesis (Chapter 9).

Universal Design: The design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialised design (Rao, Ok and Bryant, 2014).

User-Sensitive Inclusive Design (USID): A design approach rooted in both user-centred and inclusive design, where the term ‘centred’ is substituted with ‘sensitive’ to support population diversity (Gregor, Newell and Zajicek, 2002).

Vision-to-Language (V2L) Models: AI models, often using encoder-decoder methods, which process images and translate visual information into text descriptions (Section 6.1).

Web Activity (WA): A measure referring to web activity rates per disability, used in calculating the CxDAI (Berger *et al*., 2010).

Web Content Accessibility Guidelines (WCAG): The most popular set of accessibility guidelines devised within the W3C’s Web Accessibility Initiative, used to rank web content on a conformance scale (W3C, 1997).

Chapter 1. Introduction

1.1 Background and motivation

Ensuring that web content is accessible to all is a professional responsibility of web content creators. However, recent annual accessibility reports suggest that accessibility barriers persist in 94.8% of website home pages (WebAIM, 2025), which is but a 3.3% improvement in the last five years (WebAIM, 2020). It has, in fact, been shown that web content creators are often reluctant to cater to the accessibility of web content, which is more evident for barriers where training is needed to address them appropriately (Hanley *et al.*, 2021). Two of the most pivotal such barriers are unavailable and unsuitable *alternative text (alt text)* for images on the Web. Alt text is "a textual substitute for non-text content in Web pages" (WebAIM, 2021), which is accessed via *screen readers*, i.e., assistive software that reads out loud content displayed on computer screens (Dobransky and Hargittai, 2016). Blind people are the primary users (76.6%) of screen readers and they are unable to interact with web non-text content when alt text is missing (WebAIM, 2024a). In response, automated approaches have been used to address the reluctance of web content creators to cater to the accessibility of web content and the missing alt text barrier by scaling the generation of alt text descriptions. Alt text inclusion on the Web has in fact increased by 12.8% in the last five years (WebAIM, 2025); however, unsuitable alt text on the Web has only increased by 4.1% during the same period.

Suitable alt text is defined in this thesis, as alt text that is accurate, complete and concise in relation to the *context* in which the image it substitutes is used in (Mack *et al.*, 2021), while it has been shown that unsuitable alt text can be equally or more problematic than missing alt text (Salisbury, Kamar and Morris, 2017). Alt text suitability is in fact operationalised in this work based on these three dimensions:

1. Accuracy: Does the alt text accurately reflect the role of the image in the context it is used in? (Chapter 9)
2. Completeness: Does the alt text align with user expectations based on their experiential understandings? (Chapter 8)
3. Conciseness: Does the alt text capture what is important about the image it substitutes in a verbose/concise manner as per the image's role? (Chapter 10)

Authoring alt text suitably is a difficult task, which is underpinned by a mismatch between the perceptions of web content creators and consumers' hands-on experiences with alt text (Harris, 2020). Perspectives between accessibility experts have also been shown to vary in the case of alt text suitability, especially considering that there are people who use screen readers

who are not blind and thus expect different information in alt text descriptions (Lengua, Rubano and Vitali, 2022). Central to most takes on suitability is the aforementioned need to tailor the alt text to the context in which the image it substitutes is used in; however, context has been loosely defined in past work, making it difficult to use this concept to train people in authoring alt text suitably (Miranda and Araujo, 2022). Inevitably, the need to scale alt text generation on par with the increased abundance of multimedia content on the Web points to automated approaches, wherein the recent surge in large-scale artificial intelligence (AI) models can help address scalability (Mitchell, 2021). However, AI models have fallen short of guaranteeing the suitability of alt text, and there is also a lack of models that consider context owing to gargantuan noisy datasets used for training such models, which lack contextual information (Birhane, Prabhu and Kahembwe, 2021). It is therefore necessary that automation be complemented with the collection of alt text descriptions that are context-driven and of sufficient volume to train AI models, aiming at close to human-level alt text quality whilst automating the process.

Whereas AI models for the generation of text descriptions require large-scale datasets to be trained appropriately (Changpinyo *et al.*, 2021), manual alt text authorship by experts is costly to scale and prone to the aforementioned disparity in experts' takes on suitability (Abuaddous, Jali and Basir, 2016). Therefore, crowdsourcing approaches handing alt text authorship to large non-expert crowds have been used; however, only a few such approaches have been reported in the literature, and neither incorporate training of non-experts nor consider context in alt text, which is vital given the difficulty of authoring alt text suitably (Gleason *et al.*, 2020). Crowdsourcing literature in fact suggests the game-based crowdsourcing approach, i.e., *Games-With-A-Purpose (GWAPs)* for more difficult tasks and when scalability is paramount (Aliady and Poesio, 2024). The benefits of GWAPs over other crowdsourcing approaches are tied to games excelling at training users (Tuite, 2014) and motivating participation via non-monetary incentives (gameplay enjoyment) (Chamberlain *et al.*, 2013), respectively. Despite these benefits, and GWAPs having yielded promising results for similarly complex tasks (e.g., Madge *et al.*, 2022; Lafourcade and Le Brun, 2023), there is **no reported** GWAP for context-driven alt text annotation.

1.2 Research questions

To achieve the aims discussed in the previous section, a set of overarching research questions (RQs) were defined, underpinning the research conducted in this thesis:

RQ1. What are the main web accessibility barriers and what is their impact across different user groups of people with disabilities?

There is no recent work on reviewing the web accessibility landscape including accessibility barriers and diverse disabilities. There have been past similar efforts (e.g., van der Smissen *et al.*, 2020; Acosta-Vargas *et al.*, 2021), but they were largely restricted to specific sectors and disabilities. To answer RQ1, a scoping review of the web accessibility landscape is conducted (Chapter 2), and a framework is proposed to assess the impact of accessibility barriers across disabilities (Chapter 3).

RQ2. How do the perspectives of visually impaired users and web content creators compare regarding web accessibility, barriers, and alt text suitability?

Whereas novel solutions to alt text barriers, which can complement automated approaches are necessitated, it is important to inform such solutions with the perspectives of web content creators and the experiential understandings of visually impaired users. This is very relevant in the case of alt text barriers, where a mismatch in their perspectives has been reported (Harris, 2020). To answer RQ2, the following S-RQs will need to be answered through a user study that will involve semi-structured interviews with both visually impaired users and web content creators (Chapter 8):

- S-RQ1. What are the perceptions of web content creators on the accessibility of the web through screen readers against visually impaired users' web navigation experiences?
- S-RQ2. What are the perceptions of web content creators on WCAG against those of visually impaired users?
- S-RQ3. What makes alt text suitable according to both visually impaired users and web content creators?

RQ3. Is a GWAP an efficient approach to the generation of human-centred, context-driven alt text at scale?

The need to gather human-centred datasets to train AI models for automatically generating alt text is a pending issue, due to the poor quality of generated alt text and the small scale of the datasets. To address the poor quality of alt text, the context in which the images are used in (see section 1.1) and the recommendations from users (RQ3) are utilised to train non-experts in authoring alt text. To evaluate the scalability of the GWAP approach, non-expert players are

recruited to play the GWAP developed in this thesis as part of a user study (Chapter 10). This study will need to answer the following S-RQs that will help address RQ3:

- S-RQ4. How effective is the implemented solution in generating alt text descriptions at scale?
- S-RQ5. How does the use of structured context prompts influence the quality of player-authored alt text?
- S-RQ6. How does the descriptive verbosity of player-authored alt text influence its quality perception?
- S-RQ7. What levels of consistency or divergence emerge among player ratings?

RQ4. Is it possible to generate human-level quality alt text descriptions whilst automating alt text generation through AI?

This final RQ focuses on the final output of the AI models trained on the GWAP-generated dataset (Chapter 11). Whilst automation is vital for scalability, current approaches fail in terms of alt text suitability. Evidently, to answer RQ4, all of the above RQs, which underpin the reviews of relevant literature and user studies, need to be answered first. Further, the evaluation of the performance of the models in terms of alt text quality and ability to generate context-driven alt text will help answer the following S-RQs, which are necessary to address RQ4:

- S-RQ8: Is the use of a GWAP-generated dataset to train the AI model (trained model) for generating alt text descriptions an effective approach to get closer to a human average compared to pure image processing (control model)?
- S-RQ9: How well does the trained AI model classify decorative (eye candy) images?
- S-RQ10: To what extent does learning from structured context prompts improve context-driven alt text generation?

1.3 Aims and objectives

To answer the above RQs, the appropriate research methodology and philosophy underpinning it needed to be chosen. In this vein, pragmatist philosophy was chosen (Chapter 7), owing to the need for a practical, deployable solution, in this case, the GWAP developed in this thesis (Chapter 9). Unlike alternatives, such as interpretivism and positivism, pragmatism allows for the use of a mixed-methods approach, which was deemed necessary due to accessibility studies pointing towards the need to also gather empirical insights from screen reader users (Chapter

8). Qualitative insights do not map well to positivism which seeks objective measures, while interpretivism does not map well to the need for a practical solution, which was based on the analysis of the insights deriving from both alt text authors and consumers. Accordingly, the aim of this research is first to investigate the effectiveness of a crowdsourcing game-based approach to gather a human-curated dataset of context-driven alt text descriptions and then to evaluate the effectiveness of the dataset for training AI models to automatically generate alt text that is context-driven and close to average human-level quality. A GWAP will thus be designed and developed as part of this thesis to address the dual challenge of inadequate authoring practices and alt text unsuitability. The human-curated dataset deriving from the GWAP will be used to train AI models to automatically generate alt text descriptions, with the objective of addressing the dual challenge of authorship reluctance and scalability. Importantly, it is clarified that this thesis makes **no** contribution to the AI field; rather, existing AI models are used as tools to contribute to the field of Human-Computer Interaction (HCI). Accordingly, the main objectives of this research are to:

- OBJ1: Provide a review of the state of the web accessibility landscape with the objective of understanding barriers and their impact across disabilities.
- OBJ2: Provide a review of alt text barriers and state-of-the-art solutions.
- OBJ3: Investigate the contrasting perspectives of web content creators and consumers on the suitability of alt text.
- OBJ4: Develop and propose a GWAP for generating suitable and context-driven alt text based on the investigation of web content creator-consumer perspectives.
- OBJ5: Evaluate the effectiveness, quality, and consistency of the GWAP approach for alt text annotation and evaluation.
- OBJ6: Train AI models for automated alt text generation on the GWAP-generated dataset, with the objective of comparing their performance with pure image processing in terms of approximating human-level quality alt text.
- OBJ7: Evaluate the performance of the models both in terms of the quality and the presence of context in automatically generated alt text descriptions.

For completeness, Fig. 1 below reveals the hierarchical relationship between the research aim, RQs (RQ1-RQ4), and objectives (OBJ1-OBJ7) by showing how each RQ stems from the research aim, as well as what objectives are connected to each RQ.

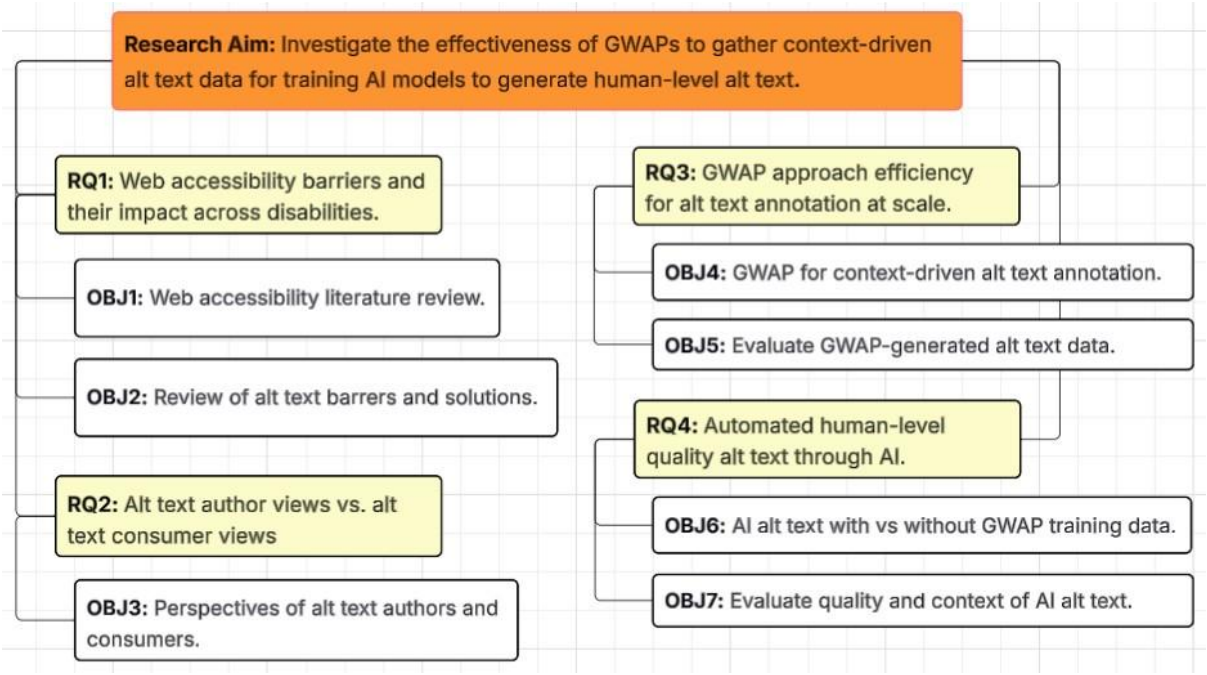


Fig. 1. Diagram connecting RQs to research aims and objectives

1.4 Contributions

The main contributions of this thesis are as follows:

- An Impact Assessment Framework, which is the first, to the best of the researcher's knowledge, attempt to measure the impact of web accessibility barriers by taking into account disability-specific considerations (Chapter 3).
- The first structured semantic definition and syntax of **Alt Text Context (altC)**, accounting for multiple factors that influence how an image should be described in alt text and which is framed within two important elements related to the image and the webpage being used in. It can serve as a framework towards improving the relevance and informativeness of alt text beyond traditional isolated image labelling (Chapter 4).
- Alt text trainability and suitability recommendations drawing on a set of six themes, which are to the best of the researcher's knowledge, one of the first such efforts to compare and bring together the views of web content creators and visually impaired users on alt text suitability (Chapter 8).
- The first, to the best of the researcher's knowledge, GWAP for alt text annotation and evaluation that embeds context by design, which is also one of the very few crowdsourcing efforts for such type of annotation (Chapter 9).
- A novel dataset of player-authored alt text descriptions and rating scores that considers

context in alt text, which is available upon reasonable request[1] (Chapter 10).

- Two proof-of-concept AI models that integrate contextual features during training and assessing their impact in shaping model performance, thereby offering empirical evidence for the impact of contextual cues on alt text generation (Chapter 11).

1.5 Structural overview

The order of the discussion in this thesis begins with the broader concepts of accessibility and disability before moving on to web accessibility barriers, highlighting the need to address alt text barriers in particular. State-of-the-art solutions to alt text barriers (incl. manual, automated, and crowdsourcing solutions) are then discussed, and the methodological theory underpinning the research work in this thesis is presented. Finally, the remaining chapters present the relevant user studies, the implemented game-based solution and the use of its output to train AI models for automated alt text generation, as well as related evaluations (see Fig. 2 below).

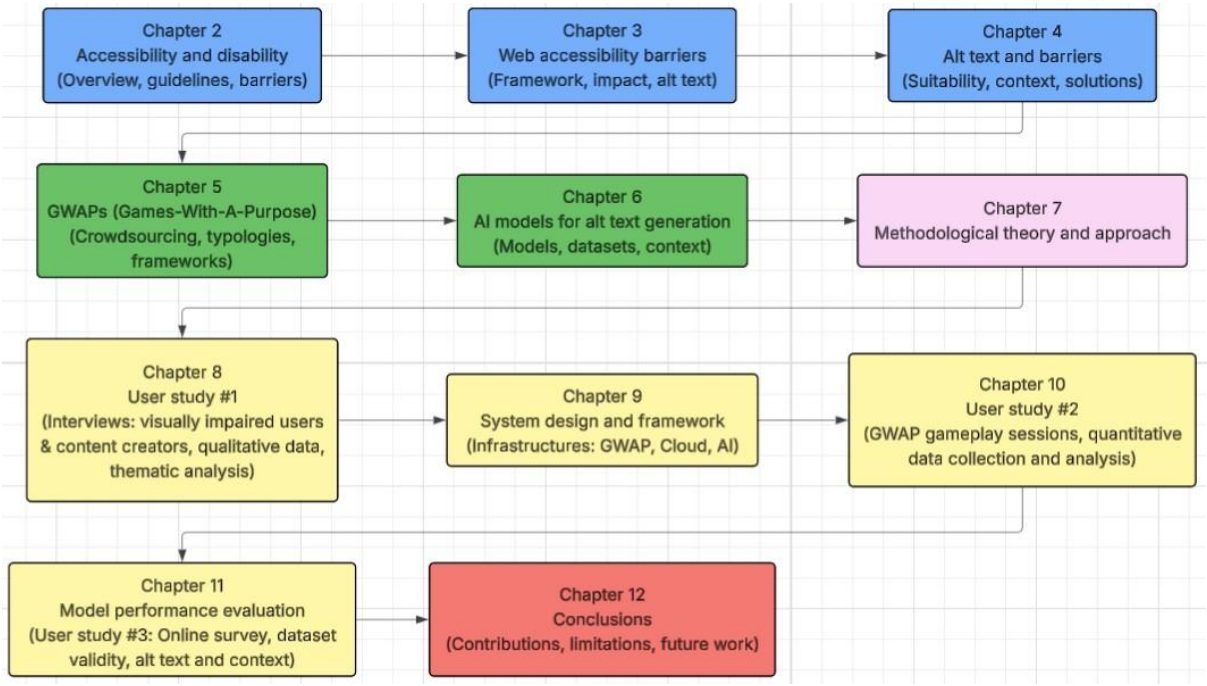


Fig. 2. Thesis structure, from accessibility background (blue) to GWAPs and AI background (green), methodology (pink), user studies and implemented solution (beige), and conclusions (red)

[1] The dataset is available from the researcher - Email: nick.droutsas@brunel.ac.uk

Chapter 2. Accessibility, disability, and the Web

2.1 Introduction

Designing for accessibility is a widely recognised practice which is underpinned by legal directives such as the European Disability Act 2019, (European Parliament and Council, 2019), the 1998 Rehabilitation Act in the USA (Section508, 2023), and the Equality Act of 2010 in the UK (Lawson, 2011). Accessible products are in fact 35% more usable by everyone and are typically cheaper to run and maintain (AbilityNet, 2017). However, a recent survey has found that 98.1% of the top 1 million home pages and 97.8% of more than 100,000 interior pages within a wide sample of websites had detectable accessibility failures (WebAIM, 2020). Despite recent acknowledgements in research on web accessibility about the importance of inclusive web design and development (Rodriguez, 2020; Vilares *et al.*, 2020; Lundgard and Satyanarayan, 2022), there is still a plurality of reasons such as dissonance in accessibility understandings (Petrie, Savva and Power, 2015), overreliance on solutions devoid of user nuances (Gartland *et al.*, 2022), insufficient resource allocation towards accessibility (Open Inclusion, 2022), reluctance to engage with training on web accessibility (Abuaddous, Jali and Basir, 2016; Williams *et al.*, 2022), and a general mismatch between the perspectives of content creators and content consumers about what makes web content accessible (Hanley *et al.*, 2021), that have contributed to the emergence and persistence of web *accessibility barriers* which hamper efforts aimed at improving web accessibility. Accordingly, this chapter presents an overview of the current state of web accessibility and related approaches and guidelines.

2.2 Disability definitions and models

Before exploring the current state of web accessibility, it is imperative that disability is first understood. Disability is part of the human condition, but many people with disabilities do not have equal access to services that are taken for granted, such as equal access to the web, and are therefore excluded from everyday life activities. Despite the importance of this issue, there is no agreed definition of disability. In response, the World Health Organization (WHO) put forward the notion that:

> "*disability results from the interaction between individuals with a health condition…with personal and environmental factors including negative attitudes, inaccessible transportation and public buildings, and limited social support*."
> (World Health Organization, 2023)

This is also the definition adopted in this thesis and it identifies three dimensions that characterise disability, which are a person's impairment (i.e., body structure or function, or

mental functioning), an activity limitation (i.e., difficulty seeing, hearing or walking) and restrictions to participation in daily activities (i.e., working, socializing, healthcare) (Rosenbaum and Stewart, 2004). There are also many types of disability, such as those affecting a person's vision, hearing, mobility, cognitive functioning, as well as psychological disability. Disability can also be invisible as it may not be immediately apparent to others. It can be therefore agreed that disability is complex, dynamic, and multidimensional.

The above definitions also highlight the role of medical and social barriers in a person's disability. There are in fact two prevailing models of disability - the medical and the social model. The former argues that "a person's functional limitations (impairments) are the root cause of any disadvantages experienced and these disadvantages can therefore only be rectified by treatment or cure," whilst the latter discusses that disabilities are a result of societal barriers rather than people's bodies (Crow, 2010, pp. 137-138). Therefore, disabilities are different to impairments and are framed as a limitation of a person in the medical model and a barrier in the social environment in the social model. Research views on these two models vary, as the social model is often considered as too utopic in practice (Shakespeare, 2006), whilst others argue that disability is indeed described as a social prejudice. Historically, both of these models have been presented as distinct; nevertheless, recent efforts by the WHO resulted in the International Classification of Functioning, Disability and Health (ICF) that instead presents that functioning and disability are interrelated and are characterised by a dynamic interaction between health, personal and environmental factors (Rosenbaum and Stewart, 2004), which is described as the 'bio-psycho-social' model (World Health Organization, 2011), and has now replaced the medical and social models as the prevailing model of disability.

2.2.1 Disability policy and legislation

In addition to the shift from the medical and social models to the 'bio-psycho-social' model of disability, legal policies and laws are also in place, giving people with disabilities important rights not to be discriminated against. One of the first laws to take effect was the *Rehabilitation Act of 1973* in the US, which prohibited organizations that had received government funding from discriminating against people with disabilities (Persson *et al.*, 2015). *Section 508* was later included within the Rehabilitation Act to cover accessibility aspects in electronics and in Information and Communication Technologies (ICTs), which was revised in January 2018 to comply with the latest accessibility requirements (Taylor and Bicak, 2019). Another pivotal law for accessibility in the US is the *American Disability Act* (ADA) (Persson *et al.*, 2015).

Similarly, in the UK the *Disability Discrimination Act* (DDA) was established in 1995 classifying discrimination as the unjustified and unfavourable treatment of certain individuals within the population (Sanderson-Mann and McCandless, 2005). The *Special Educational Needs and Disability Act* (SENDA) later extended the DDA to address discrimination in an educational context, which was followed by the *Equality Act 2010* amendment to encompass neglected indirect discrimination cases (Lawson, 2011). More recently, the *EU web Accessibility Directive*[2] and the *European Accessibility Act* (EAA)[3] were introduced in the European Union in 2016 and 2019, respectively, to focus more closely on ensuring equal access to the web and the built environment. However, it is argued that both of these EU regulations face a 'knowing-doing gap,' that is, "the gap that traditionally occurs when people 'know' what should be done but there [are] a number of challenges and barriers in place that prevent action being taken" (Marcus-Quinn, 2022, p. 6). This can also hold true for similar laws and policies in the US and the UK, as well as to the rest of the world, which calls for a much-needed review of accessibility barriers and their impact (see chapter 3).

2.3 Accessibility design approaches

The previous section highlighted that equal access for people of all abilities is supported by many legal acts and policies. In a similar vein, the academic community and the industry have put a larger focus on accessibility and, specifically, the design of accessible applications to support people with various disabilities (Newell and Gregor, 1999). Since then, a range of approaches, methods and tools have been developed to enable the design of accessible technologies (Abascal and Nicolle, 2005). In this section, accessibility and the main design approaches are discussed.

2.3.1 Defining accessibility

Historically, despite efforts to define accessibility tracing back to the 1950s, the plurality of perspectives on its essential components have delayed the development of an unambiguous conceptual framework of accessibility (Petrie, Savva and Power, 2015; Stratton *et al.*, 2022). It is however noted in scholarly work that the terms *property (*"the right to benefit from things") and *access* ("the ability to derive benefits from things") were at the heart of discussions about accessibility (Ribot and Peluso, 2003, p. 153). Traditionally, both physical and digital environments have been susceptible to the *environmental docility hypothesis*, that is, "the less

2 https://digital-strategy.ec.europa.eu/en/policies/Web-accessibility
3 https://ec.europa.eu/social/main.jsp?catId=1202

competent the individual, the greater the impact of environmental factors on that individual" (Lawton, 1986, cited in Iwarsson and Ståhl, 2003, p. 59), which is lacking encompassment of *population diversity* (Lange and Becerra, 2007; Hosking, Waller and Clarkson, 2010) that is evident in disability. More recently, it has been argued that modern definitions of accessibility should recognise the concept of *population diversity*, which considers a multitude of abilities and contexts (Waller *et al.*, 2015; Heylighen, Van der Linden and Van Steenwinkel, 2017).

It is, therefore, important to consider alternative perspectives on accessibility, particularly those that aim for the recognition of a diverse pool of activities and their interplay within the environment. Such perspectives should encompass population diversity and should recognise the interrelatedness between accessibility and disability (Persson *et al.*, 2015). It can be hence conjectured that a contemporary unified definition of accessibility may only emerge through acknowledging such components alongside digital advancements, which can entail multiple benefits for deepening the discourse around novel and proper accessibility practices (Petrie, Savva and Power, 2015; Stratton *et al.*, 2022). Despite scholarly perspectives on accessibility being disparate, there is now great acquaintance with the term accessibility (Iwarsson and Ståhl, 2003), and the related design approach of *accessible design* which relies on extending standard design principles to people with some type of functioning limitation (ISO and Guide, 2001). In a similar fashion, *barrier-free design* is a related design approach focusing mainly on lifting barriers for specific individuals in order to perceive an environment as barrier-free, which is arguably considered a highly-subjective notion (Persson *et al.*, 2015). It is notable that in both approaches there is, again, an overreliance on the environment, and they both seem to fail to acknowledge population diversity, which should be key in modern definitions.

2.3.2 Usability and design approaches

Previous work by Carlsson, Iwarsson and Ståhl (2002) suggests that accessibility should be anchored to thorough knowledge of population diversity, which should be a foundational component of any contemporary definition of accessibility. However, accessibility seems to lack what Iwarsson and Ståhl (2003, p. 62) note as the *activity component*, i.e., a "description of activities to be performed by the individual or group at target, in the given environment." In this vein, accessibility is preconditioned by usability, which allows for extending beyond a specific disability. The close connection between accessibility and usability has been studied before (Aizpurua, Harper and Vigo, 2016; Bi *et al.*, 2022).

The need for usability irrespective of disability or people's abilities resulted in the *universal design approach*; i.e., "the design of products and environments to be usable by all people, to

the greatest extent possible, without the need for adaptation or specialised design" (Rao, Ok and Bryant, 2014), which is closely connected to population diversity, with its number one goal being *universal access* to avoid discriminatory design (Imrie, 2012). Relatedly, the term *design for all* has been used more popularly in Europe to describe universal design approaches (EEID, 2004). Nevertheless, such approaches accept that it is impossible to always adequately and irrespectively represent the entire population through design ((Waller *et al.*, 2015). Similarly, *inclusive design* is another popular design approach, which has its roots in product design (Heylighen, Van der Linden and Van Steenwinkel, 2017). The term has grown in significance, particularly across the UK, and it is described as an attitude to design that seeks to evolve based on continuous gathering of insights about population diversity, diverse experiences, and interactions within the environment (Persson *et al.*, 2015).

Whilst all aforementioned design approaches fundamentally share the same goal, inclusive design is the only one that considers the flexibility and ability of the designer to loosen design efforts aimed at universal access, social inclusion and design resonance if those are deemed prohibitively costly or hard to implement (Scott, Spyridonis and Ghinea, 2015). Therefore, there is a need to shift from adjusting a design to accommodate specialised needs to shaping the environment in a way that it is independent of accessibility and disability. Accordingly, user-centred design (UCD) approaches which take into account the variety of ways a user, regardless of ability or preferred modes of functioning, may possibly interact with the environment have been considered (Droutsas, 2021). For accessibility, Persson *et al.* (2015) argue that *User-Sensitive Inclusive Design* (USID) is a more prominent approach, as it is rooted in both user-centred and inclusive design. In USID, the term 'centred' is substituted with the term 'sensitive' to support population diversity, and the term 'inclusive' is added to encompass the flexibility and ability of the designer inherent in an inclusive design approach (Gregor, Newell and Zajicek, 2002). Whereas an abundance of attitudes to designing for accessibility, usability, inclusivity and utility have been suggested, Stratton *et al.* (2022) propose the need to reimagine, and possibly also repurpose, accessibility design approaches to suit more modern digital settings, such as the web, video games, etc.

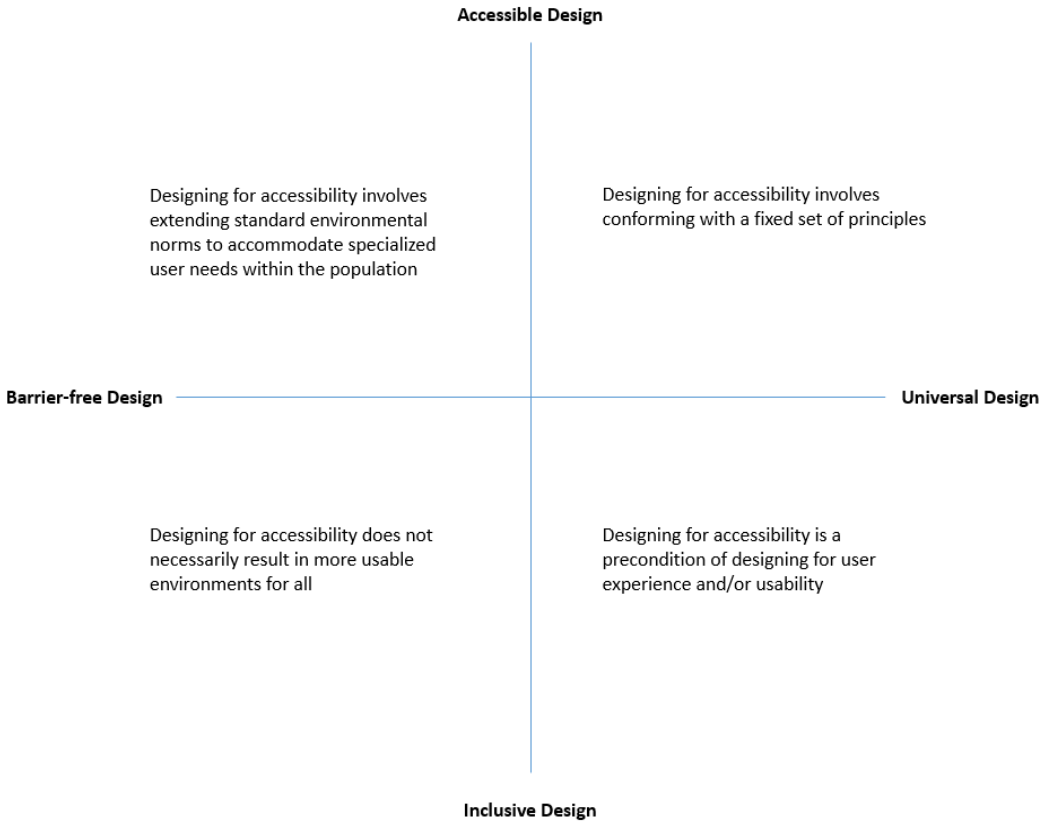


Fig. 3. Classification matrix of accessibility design approaches inspired by Avgerou (2010)

The various perspectives and their interplay are summarised and classified in the above figure, which maps the relationships between the discussed design approaches on a matrix, showing how distantly- or closely-related such approaches are; therefore, highlighting the nuanced and dynamic nature of designing for and delivering accessibility. For clarity, UCD and design for all are not included in the figure, as the former is not accessibility-specific and the distinctive qualities of design for all are identical to those of universal design. The matrix also highlights a **mentality gap** in design efforts towards accessibility, which aligns with the two prevailing models of disability (see section 2.2). Further, this gap feeds into a functional gap between the number of ways users want to interact with an environment and the actual number of ways that are offered/denied to them in practice (Aizpurua, Harper and Vigo, 2016). Evidently, the fact that advancements in accessibility have been largely outpaced by advancements in technology stresses the importance of this gap in digital contexts (Stratton *et al.*, 2022). Accessibility of the Web in particular is a pending issue, where rapid technological advancements and dissonant design decisions have immediate impact on accessibility, usability, and user experience. It is therefore necessary to inquire into accessibility understandings, practices, and barriers, as those emerge on the Web, to understand their impact on users and propose appropriate solutions.

2.3.3 Web accessibility

Accordingly, web accessibility is defined as the ability of:

> "*all people, particularly disabled and older people, to use websites in a range of contexts of use, including mainstream and assistive technologies; to achieve this, websites need to be designed and developed to support usability across these contexts*." (Petrie, Savva and Power, 2015, p. 3)

The above definition is chosen in this work, as it agrees well with inclusive design approaches. Additionally, the definition considers various reported inconsistencies in accessibility-related terminology. This promotes clarity for setting consumer demands and provides a common framework for discussing issues, past approaches, and solutions. However, web accessibility has been treated by industry as a low priority consideration that is traditionally limited to compliance with a set of official guidelines and legal mandates (Miranda and Araujo, 2022).

2.3.4 Web accessibility guidelines

Central to scholarly work and practices are the *Web Content Accessibility Guidelines* (WCAG), which were devised within the W3C's Web Accessibility Initiative (WAI), and are the most popular accessibility guidelines (W3C, 1997). Since their development back in 1999 (WCAG 1.0), there have been three more iterations (WCAG 2.0, 2.1 and 2.2) with their latest one (WCAG 2.2) pushed out in October 2023 (W3C, 2023). The WCAG ranks web content on accessibility per a hierarchically arranged scale of three levels - Level A, AA and AAA - with level A being the minimum and level AAA being the maximum level of conformance with the WCAG (W3C, 2023b). The emergence and wide acceptance of WCAG has helped provide a common blueprint for content creators of varying abilities (Kercher and EIDD, 2008; Lengua, Rubano and Vitali, 2022), and WCAG conformance is an ironclad first step for making websites accessible (Cooper, 2016; Dobransky and Hargittai, 2016). The complexity of understanding and applying the WCAG, however, has been reported (Spyridonis, Daylamani-Zad and Paraskevopoulos, 2017; Spyridonis and Daylamani-Zad, 2019, 2021), which inevitably has led to inadequate WCAG conformance by a notable number of websites (Crespo, Espada and Burgos, 2016).

Foremost, there is now sufficient evidence to suggest that conformance to standards and guidelines alone is by no means a complete picture of web accessibility (Iniesto *et al.*, 2022; Vollenwyder *et al.*, 2023). This pertains to conformance neglecting the previously mentioned population diversity, with Aizpurua, Harper and Vigo (2016) separating web accessibility per

conformance with guidelines from accessibility per users' hands-on experience. For example, Vollenwyder *et al.* (2023) showed how websites conforming WCAG Level AA continued to present significant accessibility barriers to diverse users, owing to WCAG's focus on technical compliance, rather than how differently each barrier can affect the web navigation experience of diverse users with impairments.

Relatedly, the *British Broadcasting Corporation* (BBC) guidelines also emphasise how web accessibility is a continuous strive rather than an act of conformance, and as such include information on who will benefit, as well as how will they benefit, from the incorporation of each accessibility feature (BBC, 2023c). The BBC guidelines are less complex than WCAG and also aim to foster empathy towards inaccessible web navigation experiences, but they are not as thorough. An in-depth understanding of accessibility through diverse user perspectives is thus imperative to inclusively embed accessibility. However, past research has intimated a perception mismatch between web content creators (i.e., individuals responsible for authoring web content) and web content consumers (i.e., individuals who interact with web content) in relation to what makes web content accessible (Harris, 2020). This mismatch calls for qualitative research efforts to help bridge such perspectives and propose recommendations for practice (see Chapter 8). It is therefore evident that the accessibility of websites cannot be guaranteed with conformance with web accessibility guidelines alone (Power *et al.*, 2012; Iniesto *et al.*, 2022), as they fail to encompass population diversity, which is essential to understand the diverse ways different accessibility barriers impact different users with disabilities on the Web.

2.4 Web accessibility barriers

Accessibility barriers refer to "*any obstacles that prevent individual, especially those with disabilities, from accessing or interacting with online content effectively*" (Halpin, 2025, para. 2). It is thus evident that accessibility barriers can have various forms, with insufficient web accessibility knowledge being largely considered the most important barrier (Nedelkina, 2022). Past work has shown that limited time to invest in accessibility (Lengua, Rubano and Vitali, 2022), and low motivation to engage with training (Bi *et al.*, 2022) are common causes of the insufficient knowledge barrier. Further barriers include the difficulty to find and recruit web accessibility experts (Abuaddous, Jali and Basir, 2016), poor communication of accessibility benefits to content creators (Open Inclusion, 2022), and the lack of involvement of content consumers in accessibility decisions and processes (Vollenwyder *et al.*, 2020).

Whilst the above barriers relate to human roles within web accessibility, barriers can also be found in tools, such as automated authoring and evaluation tools, which support the process of producing accessible products and help detect accessibility issues in web content, respectively (Chisholm and Henry, 2005; Kaur and Kumar, 2015b). The key barrier in such tools is that they are typically limited to WCAG compliance and lack consideration of population diversity (Moreno *et al.*, 2019). Assistive technologies, on the other hand, which include hardware and software tools typically used by users with disabilities to overcome barriers while navigating the Web (Amado-Salvatierra *et al.*, 2016) are not adequately available for all types of disability (Ismailova and Inal, 2022). In hindsight, it is important to emphasise the complementary roles of humans and tools to ensure sustainable delivery of accessible web content.

Web content has in fact been growing increasingly more complex with the incorporation of novel technologies and multimedia into web design (Stratton *et al.*, 2022). Results of WebAIM (2025)'s recent web accessibility report reflect this, with 94.8% of website home pages failing to comply with WCAG (see section 1.1). Further, WebAIM has used six categories of barriers claiming that these cover 96% of all potential accessibility barriers that surface on the Web, i.e., *low contrast text, missing alternative text for images, empty links, missing form input labels, empty buttons,* and *missing document language*. In the case of low contrast text, Friedman and Bryen (2007) documented that text clarity and the use of color for contrast have been frequently cited in web accessibility guidelines for users with cognitive impairments. Similarly, Brewer (2011) discussed how users with dyslexia are affected by low contrast text, as it renders web content comprehensibility more difficult. This is also appreciated by McCarthy and Swierenga (2010), who advised implementing text configurability to make websites more accessible to these users. Relatedly, visually impaired users, as the primary users of screen magnifiers (Zong *et al.*, 2022), can also be impacted by barriers like low contrast text, and poor or no text contrast between the background and foreground (Ruth-Janneck, 2011).

Missing alt text for images has also been shown to hamper the web navigation experience of blind users and visually impaired users who are otherwise unable to discern non-textual web content (Power and Petrie, 2007; Kenigsberg *et al.*, 2019); secondarily, missing alt text can also hamper the experience of users with cognitive impairments or dyslexia, as well as users with motor disabilities who have also been shown to opt for using screen readers to smoothen their web interactive experiences (Vollenwyder *et al.*, 2023). Earlier research by Takagi *et al.* (2009) reported that 124 out of 323 users were disturbed by the unavailability of alt text for images, agreeing well with McEwan and Weerts (2007)'s findings that missing alt text is the most pivotal accessibility barrier upon compilation of a diverse range of accessibility studies.

Interaction issues during web browsing such as empty links, missing form input labels, and empty buttons are directly linked to experiential aspects such as prejudices, evoked memories, and expectations, and thus they have been shown to cause frustration to users with cognitive impairments (Campoverde-Molina, Luján-Mora and Valverde, 2021), visual impairments (Kaur and Kumar, 2015a), and blind users (Aizpurua, Harper and Vigo, 2016). Frustration due to the use of such seemingly operable interaction options in websites can also be experienced by users with dyslexia (Vázquez and Torres-del-Rey, 2019), and users with motor disabilities due to low or no alternative input device operability (Ruth-Janneck, 2011). It is thus evident that barriers can take various forms and their impact across the diversity of disabilities can also vary, which calls for practical solutions to assess their impact per disability (Chapter 3).

2.5 Chapter summary

In this chapter, the key definitions and models of disability, as well as key policies, regulations, accessibility design approaches, and web accessibility barriers were discussed. This allowed for an appreciation of the current state of web accessibility, which is largely underpinned by conformance with guidelines and standards, such as the WCAG and the BBC guidelines. It was further discussed that WCAG conformance is an ironclad first step towards the accessibility of websites, but it is not the full picture of web accessibility. Despite this, most web accessibility efforts typically start and end with conformance to standards and guidelines (Kaur and Kumar, 2015a, 2015b; Moreno *et al.*, 2019). Consequently, persistent barriers dominate the past and current state of web content accessibility (Vollenwyder *et al.*, 2023). However, there is a gap in scholarly knowledge on the impact of accessibility barriers on the web navigation experience of users with diverse disabilities (Miranda and Araujo, 2022). Accordingly, Chapter 3 presents and proposes a framework for measuring the impact of barriers across disabilities.

Chapter 3. Barrier impact assessment framework

3.1 Introduction

As discussed in the previous chapter, relying solely on conformance fails to address population diversity and consumers' expectations and needs, which is particularly important for users with disabilities (Berger *et al.*, 2010). Whilst barriers, such as content creators' lack of knowledge in accessibility and overreliance on WCAG conformance, can impact accessibility, it is web content where interaction can be distinct per disability, and whose impact must be measured by considering population diversity. Past research (Persson *et al.*, 2015; Vollenwyder *et al.*, 2023) has produced a much-needed review of measures reflecting web accessibility barriers' impact by taking into account disability-specific considerations. However, it has been intimated that existent measures are most often limited to one disability (Morris *et al.*, 2018; Muehlbradt and Kane, 2022), or are reporting the general prevalence of each barrier on the web (WebAIM, 2020, 2025). To the best of the researcher's knowledge, the impact of web content accessibility barriers per disability has not yet been investigated and reported, despite past research in accessibility (Friedman and Bryen, 2007; Dobransky and Hargittai, 2016) having signalled the increased need for such an investigation. Accordingly, this chapter presents the attempt to gather and amalgamate disability-specific insights to approximate the impact of web content accessibility barriers across disabilities. First, the measures that were used to calculate said impact (Section 3.2) are described; then, these measures are leveraged to explain how the proposed assessment framework for the impact of barriers across disabilities came to be (Section 3.3); and, finally, recommended use cases of the framework are provided (Section 3.4).

3.2 Web activity (WA) and disturbance rate (DR)

The web activity (WA) and disturbance rate (DR) measures per disability are inspired by Berger *et al.* (2010)'s early report on users' disturbance and web presence rates. Specifically, in Table 1 the DR per disability based on a range of recent web accessibility studies (Yeratziotis and Zaphiris, 2018; Bernard, 2019; Sala *et al.*, 2020; Noble, 2021; Rivero-Contreras, Engelhardt and Saldaña, 2021; Griffith, Wentz and Lazar, 2022) is calculated. Similarly, the term WA is used in Table 1 to refer to web activity rates. The global WA rates per disability were estimated by making use of available rates for the general population in the US (Petrosyan, 2023), Sweden (Tunberg, 2022), and Colombia (Bianchi, 2023), as well as disability-specific WA rates in Sweden (Johansson, Gulliksen and Gustavsson, 2021), which is to the best of the researcher's knowledge the only study that treats disabilities

heterogeneously. This synthesis of sources is motivated by the need to arrive at worldwide estimates of WA per disability, as WA reports are typically country-specific and treat disability as a homogeneous group. Also, to the best of the researcher's knowledge, there are currently no such available data, so estimating the above measures will allow for measuring the impact of web content accessibility barriers per disability, which is not only a significant contribution to scholarly knowledge on web accessibility, but also a practical tool for informing future web accessibility studies.

Therefore, the proposed impact assessment framework for arriving at an estimate of worldwide WA per disability begins by using these measures. Specifically, according to the most recent digital competitiveness rates per country (Taylor, 2022), Sweden which is ranked 3rd from among 63 countries, the US which is ranked 2nd, and Colombia which is ranked 60th are chosen to extrapolate the disability-specific rates worldwide. For clarity, these countries were chosen for being the highest and lowest ranked countries, respectively, in the previously mentioned digital competitiveness ranking, for which there were available data on WA for people with disabilities. Indicatively, *Formula (1)* shows how WA rates for people with disabilities in the US (63.8%) (Office of Disability Employment Policy, 2022), Sweden (80%) (Tunberg, 2022), and Colombia (34.6%) (Laverde, 2021), are applied to the disability level, and then averaged to arrive at worldwide estimates per disability, where Worldwide Web Activity (WWA) estimate, Web Activity (WA), Users with Disabilities (UD) and Users with Disabilities' Web Activity (UDWA). The results are presented in Table 1.

$$For\ each\ Country\ i\ and\ Disability\ d$$

$$WWA_d = \sum_i^n \frac{UD_i \times WA_d}{UDWA} \qquad \textbf{(1)}$$

Deferring to Formula (1), the synthesis of WA data towards worldwide estimates was needed, due to the scarcity of such statistics per country and/or disability (Johansson, Gulliksen and Gustavsson, 2021). Worldwide estimates require triangulation across countries to also consider digital divide variations between these countries; therefore, countries were selected based on the same recent digital competitiveness ranking (Taylor, 2022) to ensure representation across development spectra, whilst limitations of the approach are detailed in Section 12.5.

3.3 Cross-disability accessibility impact (CxDAI)

The WA measures are used alongside the *DR* rates calculated from (Yeratziotis and Zaphiris, 2018; Bernard, 2019; Sala *et al.*, 2020; Noble, 2021; Rivero-Contreras, Engelhardt and

Saldaña, 2021; Griffith, Wentz and Lazar, 2022), and the general prevalence of each identified barrier for web content found in the latest report from WebAIM (2020), as well as from accumulated information about whether these barriers primarily, secondarily or not at all affect each group of people with disabilities found in several studies (Keates *et al.*, 2000; Hackett, Parmanto and Zeng, 2003; Friedman and Bryen, 2007; Berger *et al.*, 2010; McCarthy and Swierenga, 2010; Brewer, 2011; Ruth-Janneck, 2011; Pascual, Ribera and Granollers, 2014; Waller *et al.*, 2015; Abou-Zahra, Brewer and Cooper, 2018; Moreno *et al.*, 2019), in order to calculate the *cross-disability accessibility impact (CxDAI)* of each web content accessibility barrier. Therefore, the CxDAI is calculated by averaging the DR and WA, as shown in *Formula (2),* where Website Homepages (HomP) and Interior Pages (IntP) and Affect Weight (AW). In this calculation, AW is proposed to represent the weight of affect for when a user is primarily (100%), secondarily (50%), or not at all (0), affected by a barrier. It has to be noted that the terminology used in the literature for each identified web content barrier differs. In this work, the terminology used is consistent with WebAIM's barrier types (WebAIM, 2023), as their annual accessibility reports are highly-acclaimed within accessibility studies.

For each Barrier i and Disability d

$$HomP_{i,j} = \{h \mid h = \ HomePage\ of\ Website_j \ \wedge \ h \xrightarrow{includes} Barrier_i\}$$

$$IntP_{i,j} = \ \{p \mid p \ \in \ Interior\ Pages\ of\ Website_j \ \wedge \ p \xrightarrow{includes} Barrier_i\}$$

$$Prevalence_i = \overline{\sum_{J}^{n} \left(HomP_{i,J} + \ IntP_{i,J}\right)}$$

$$AW_d = \begin{cases} primary \xrightarrow{is} 1 \\ secondary \xrightarrow{is} 0.5 \\ not\ at\ all \xrightarrow{is} 0 \end{cases}$$

$$CxDAI_i = \ \sum_i^n Prevalence_i \ \times (DR_d \times \ AW_d) \times WWA_d \quad \textbf{(2)}$$

For example, the CxDAI of missing alt text for images is calculated using (2) as follows:

General Prevalence of Missing Alt Text for Images: (66% homepages + 61.9% interior pages) / 2 = 63.95% webpages

Blindness: 63.95% * (80.7% * 1) * 56.3% = 29.1%

Low Vision: 63.95% * (42.6% * 1) * 61.3% = 16.7%

Hard of Hearing: 63.95% * (35% * 0) * 58.1% = 0

Cognitive Impairment: 63.95% * (40% * 0.5) * 48.2% = 6.2%

Dyslexia: 63.95% * (25% * 0.5) * 64.2% = 5.1%

Motor Disability: 63.95% * (43.5% * 0.5) * 53% = 7.4%

CxDAI of Missing Alt Text for Images: 29.1% + 16.7% + 0 + 6.2% + 5.1% + 7.4% = 64.5%

The remaining CxDAI calculations for each web content accessibility barrier in Table 1 are cumulatively calculated in the exact same fashion.

Table 1. Amalgam of primary (Blue) and secondary (Orange) web content accessibility barriers and their impact on accessibility per disability

Barrier	General Prevalence %	Blindness		Low Vision		Hard of Hearing		Cognitive Impairment		Dyslexia		Motor Disability		Cross-Disability Accessibility Impact %
		DR %	WA %	DR %	WA %	DR %	WA %	DR %	WA %	DR %	WA %	DR %	WA %	
		80.7 [a]	56.3	42.6 [b]	61.3	35[c]	58.1	40 [d]	48.2	25 [e]	64.2	43.5 [f]	53	
Low Contrast Text	85.85													64
Missing Alt Text for Images	63.95													64.5
Empty Links	61.65													69.1
Missing Form Input Labels	54.95													71.4
Empty Buttons	32.7													42.4
Missing Document Language	27.15													25.1

[a] (Griffith, Wentz and Lazar, 2022); [b] (Sala *et al.*, 2020); [c] (Yeratziotis and Zaphiris, 2018); [d] (Bernard, 2019); [e] (Rivero-Contreras, Engelhardt and Saldaña, 2021); [f] (Noble, 2021)

WebAIM (2023) discusses that these barriers account for 96.3% of web content accessibility barriers, while the absence of barriers encountered by people who are hard of hearing can be directly observed in the above table. Previous studies have, however, examined the browsing preferences and access barriers faced by people who are hard of hearing (Ruth-Janneck, 2011; Pascual, Ribera and Granollers, 2014), which include strong expectations about the presence of non-complex accompanying metadata, such as captions, subtitles and/or sign language in web audio and video content; their WA and DR rates are shown in the table as 58.1% and 35%, respectively. It is also acknowledged that the proposed CxDAI measures derive from different studies and that the taxonomy is disability-specific. A further limitation is the scarcity of research work on the impact of barriers for a range of disabilities rather than an extensive focus on blindness-related barriers (Petrie, Weber and Fisher, 2005; Miranda and Araujo, 2022), and the lack of disability-specific WA statistics per country or region, which contribute to the inherent high variability of the CxDAI measures. Deferring to the above table, barriers such as missing alt text for images are lacking in consideration of missing alt text for non-text elements beyond images (e.g., infographics), which Muehlbradt and Kane (2022) contend are more challenging to author suitable alt text for.

Admittedly, the table is only concerned with the availability of alt text descriptions, but it is disregarding the suitability of such descriptions, which has been previously framed as equally or even more problematic than alt text unavailability (Salisbury, Kamar and Morris, 2017; Mack *et al.*, 2021). If users with cognitive impairments or dyslexia were to be also considered as primarily affected by missing alt text for images, then this barrier's CxDAI would increase to 75.8% (see Formula 2), agreeing with earlier studies (McEwan and Weerts, 2007; Takagi *et al.*, 2009) highlighting unavailable and unsuitable alt text as the most prevalent barriers.

However, unsuitable alt text is not addressed in WebAIM's barrier categories, but past work has shown that it can be even more challenging than missing alt text (Salisbury, Kamar and Morris, 2017), and unlike the latter, where alt text inclusion has increased by 12.8% in the last five years, suitable alt text on the web has in fact decreased by 4.1% during the same period (WebAIM, 2020, 2025). It is therefore imperative to more holistically explore alt text barriers to better understand pending issues therein instead of sticking to alt text inclusion, as including unsuitable alt text can impose further challenges for users with disabilities. Overall, the results presented in Table 1 are, to the best of the researcher's knowledge, the first scholarly attempt to approximate the CxDAI of web content accessibility barriers using a compilation of existing knowledge on the general prevalence of such barriers and user-reported disability-specific considerations, such as DR and WA.

Whilst the measures presented in Table 1 come from diverse scholarly sources (see Table 1 footnote), representing data that were collected through different methods and across different time periods, it is considered necessary to address data uncertainty. Robustness-wise, the use of the AW which included variations (0, 0.5, 1) based on whether a user group was primarily, secondarily or not at all affected by a barrier allowed for a sensitivity analysis in the calculation of the CxDAI (see Formula 2). The presented CxDAI measures, thus, should be interpreted as illustrative orderings that could be enriched for robustness with the inclusion of further studies and statistics reporting on the DR and WA that capture barrier impact per disability.

3.4 Impact assessment framework in practice

Going forward, two main implications for practice are identified. It is first proposed that the impact assessment framework could advance automated evaluation tools (i.e., tools that detect barriers in web content) to prioritise barrier detection based on the calculated CxDAI for each barrier. This will allow such tools to address population diversity, which has been shown earlier to be problematic to address via WCAG conformance (Lengua, Rubano and Vitali, 2022). The proposed framework considers each disability independently and non-homogeneously, and can thus foster changes in future iterations of web accessibility standards and guidelines so that they better reflect each disability. Past work, in fact, hints at the need for such an improvement (Griffith, Wentz and Lazar, 2022), which maps well to the CxDAI owing to the transparency of the calculations allowing for these measures to be adjusted for each diverse audience.

3.5 Chapter summary

Building on the review of web accessibility barriers that pose considerable challenges for both web content creators and consumers, this chapter responded to the lack of efforts determining the impact of each barriers across disabilities. Accordingly and in line with past research (*e.g.,* (Friedman and Bryen, 2007; Vollenwyder *et al.*, 2023), it was argued that there should be a much-needed quantifiable measure of each barrier's impact across the diversity of disabilities, which was proposed and presented in the form of CxDAI (Section 3.3). It was further discussed that having such a measure will help move efforts away from overreliance on WCAG and towards inclusion of population diversity to address persistent accessibility barriers, such as missing and unsuitable alt text. The CxDAI framework highlighted widely reported issues with missing alt text and low contrast text even within the most fundamental web content, while alt text barriers in particular have yet to be holistically explored. In effect, the CxDAI draws on WebAIM's barrier categories, which neglect the suitability of alt text that has otherwise been

shown to pose challenges in addition to alt text availability (Mack *et al.*, 2021), with negligible recent improvements (WebAIM, 2025). Accordingly, Chapter 4 synthesises available literature on alt text for images; it inquires into alt text suitability, which is a barrier that has been neglected in WebAIM (2025), and explores state-of-the-art solutions to these barriers.

Chapter 4. Alternative text (Alt text)

4.1 Introduction

The previous chapter highlighted the prevalence of web accessibility barriers, persisting in 94.8% of website home pages (WebAIM, 2025), which is but a 3.3% improvement in the last five years (WebAIM, 2020). Alt text barriers, such as missing and unsuitable alt text, were further underscored, with the latter remaining a largely overlooked yet equally pervasive issue, as it has been reported that almost one out of three images on the Web has unsuitable alt text (Droutsas *et al.*, 2025b). Alt text is defined by the W3C as "text that is programmatically associated with non-text content or referred to from text that is programmatically associated with non-text content" (W3C, 2024b). This is typically associated with the <img> tag in HTML and it is currently only accessible via the use of screen readers, i.e., an assistive technology that reads out loud content displayed on computer screens, which is primarily used by blind people. However, a recent user survey revealed that people with other types of impairments (e.g., cognitive, hearing, motor) also report relying on screen readers (WebAIM, 2024a), agreeing well with the CxDAI framework proposed in the previous chapter.

Regrettably, alt text is often misconceived as having the same purpose as image captions (Lee *et al.*, 2023; Ramos *et al.*, 2023), which are characterised as "...echoing some information already available in the story...and... shouldn't be used to add detail that's not available in the piece", whereas alt text must "provide all the visual information available in the image" (Text descriptions working group and Cassidy, 2024). Further, past work has revealed that alt text needs not only to be present (Salisbury, Kamar and Morris, 2017), but also accurate, concise and complete in relation to the context in which the image it substitutes is used (Mack *et al.*, 2021), and to be of sufficient volume (Lee and Ashok, 2022). Relatedly, Petrie, Höckner and Rosenberger (2022) note that popular screen readers, such as JAWS and NVDA, have at times been unable to detect or grant access to alt text. Muehlbradt and Kane (2022) highlight a 22% increase in alt text inclusion from 2018 and WebAIM (2025) shows how alt text inclusion has increased by 12.8% since WebAIM (2020) for website home pages, but unsuitable alt text on the Web has increased by 4.1% over the last five years. There is also a gap as regards guidelines on what constitutes suitable alt text (McCall and Chagnon, 2022).

Central to most alt text suitability definitions is the concept of '**context** in which the image is used in'; however, disparate understandings of this hint at high variability in experts' takes on suitable alt text (Petrie *et al.*, 2011), and at a general acceptance that authoring suitable alt text is an inherently difficult and ethically fraught task (Hanley *et al.*, 2021; Mack *et al.*, 2021). Therefore, trainability, namely the ability of a solution to train individuals to author more

suitable alt text, is a further challenge (Miranda and Araujo, 2022). In order to understand the effectiveness of novel solutions to alt text barriers, it is vital to first know the scholarly discourse around alt text and the importance of alt text being context-driven for suitability. Accordingly, this chapter presents a discussion on alt text, suitability, and context, as well as a much-needed review of current alt text annotation and evaluation approaches, ranging from automated, manual, and crowdsourcing approaches.

4.2 Suitability and context of alt text

Early attempts to improve the suitability of alt text involved the use of the <longdesc> attribute in HTML, namely the use of longer descriptions that typically complement alt text when a certain character limit is reached; however, these proved to be inefficient as they often jolt-redirected the user to another webpage (O'Connell and Goldberg, 2011; McCall and Chagnon, 2022). Since then, past scholarly work has identified that such descriptions lack subject-awareness and only focus on what the image depicts (W. Chen *et al.*, 2023), while Desai *et al.* (2021) suggest leveraging the subject to allow for alt text to afford interaction with the emotion of the context the image is used in. Conversely, Srinivasan *et al.* (2021) included the page description in tandem with the image to evaluate the contextual quality of the alt text, while Mangiatordi and Lazzari (2018) argue that alt text for the same image differs based on the webpage it is used in, as well as the language the webpage is written in. However, multilingual accessibility, not least in relation to alt text, remains an untapped area (Kuppusamy and Balaji, 2023). Adding to this complexity, Zong *et al.* (2022) discuss that alt text for more complicated content such as infographics should further include information on structure and navigation beyond a description of the content itself.

Sharif *et al.* (2021), on the other hand, who conducted one of the first investigations to evaluate the accessibility of infographics, underlined how subjective the preferences of different people who use screen readers are with regard to suitability of the alt text for such infographics. Past work has, in fact, acknowledged how suitability requirements can vary per disability, as well as how conflicting such requirements can be between blind people and other people who use screen readers (Berger *et al.*, 2010; McCarthy and Swierenga, 2010; Droutsas *et al.*, 2025b). There is therefore no 'one-size-fits-all' approach to warrant alt text suitability, as it can be subjective and very much dependent on factors such as the complexity of the visual representation itself, the story piece, the webpage in which the image is used, and different screen-reader-user preferences (Crespo, Espada and Burgos, 2016; Bi *et al.*, 2022). Context,

however, has been found to be of utmost importance to suitability in past work, and it is thus argued that it should be central to any relevant efforts.

4.3 Related work on suitable alt text

In response, numerous past efforts exist that have tried to piece together what makes alt text suitable for the plurality of user needs and contexts. A comprehensive blueprint has been developed proposing that one of the first decisions is whether an image is informative (i.e., conveys information otherwise lost**)** or uninformative (i.e., either only used as eye candy or as a duplicate of neighbouring text content) (Silktide, 2020). Similarly, Silktide (2023) stressed that it is important to include alt text when it is needed, namely when it would not be repetitive of neighbouring text content or the image's caption, and when the image has a function or when it depicts something important or complex. In a related vein, Launus-Gamble (2021) proposed a syntax-based approach to what constitutes suitable alt text that is more closely tied to the image itself, which treats the image as comprising an object describing the focus of the image, an action describing what the object is doing, and a context describing where the object is doing the action. However, it must be noted that the notion of context in this syntax-based construction refers to what the image depicts, rather than the *context in which the image is used in*, as it has been previously suggested (Mack *et al.*, 2021; Bi *et al.*, 2022).

In fact, several specifications and guidelines have been put together to support producing alt text that is suitable (Droutsas *et al.*, 2025a). Practical recommendations have been proposed defining that alt text descriptions should consist of approximately 45 words and a maximum of 150 characters, be devoid of white or empty space, be written in standard language, be devoid of abbreviations, Out-of-Vocabulary (OOV) words and jargon, not only consist of numeric values or single characters, provide precise information on both content and context, maintain good flow with the rest of the content, and recognise multilingual and Unicode characters (Salisbury, Kamar and Morris, 2017; Zhong *et al.*, 2020; Mack *et al.*, 2021; Williams *et al.*, 2022). Moreover, the aforementioned WCAG (see section 2.3.4) are lacking in instructions on how to author alt text suitably. As a result, W3C complemented these with a specific technique for alt text suitability (W3C, 2023a), as well as an *alt Decision Tree* that can be used as an important guiding resource for adapting alt text per the type of image it substitutes (W3C, 2024a). The latter corroborates past work on the need for a judgment call to be made on whether an image is informative or decorative based on the content of a page and the reason for including the image on the page (Lengua, Rubano and Vitali, 2022). Along the same lines, the

BBC guidelines on alt text suitability propose an empathetic approach to describing alt text (BBC, 2023a, 2023b).

4.3.1 Context in alt text: a proposed definition

The previous section highlighted that context plays a crucial role in determining both the necessity of alt text and its suitability. This is often framed as the "*context an image is used in"*—a characterisation that acknowledges the importance of contextual factors but remains vague in its scope and definition. While this notion of context has been often reported in the literature, it lacks a clear and operationalised definition, leading to inconsistencies in how context is interpreted and applied in alt text generation which can result in limited reproduction and standardisation. This can be particularly important as we are moving towards AI-driven alt text generation which requires a structured representation of context. To the best of the researcher's knowledge, there is no context definition that addresses these challenges in the case of alt text. It is therefore necessary to move beyond the general notion of "context an image is used in" and define context in a way that is more practically applicable. Accordingly, in this work, a structured **semantic** definition of **Alt Text Context (altC)** that accounts for multiple factors that influence how an image should be described in alt text is proposed:

$$altC = f(\text{Image Type}, \text{Webpage Topic}, \text{Webpage Purpose}, \text{Image Function}, \text{Image Intent})$$

Where:

- Image Type represents the nature of the image (e.g. photograph, diagram, icon)
- Webpage Topic defines the subject matter of the webpage (e.g. climate change article, e-commerce product page)
- Webpage Purpose captures the primary goal of the webpage (e.g. informational, educational, commercial)
- Image Function describes the role of the image within the webpage (e.g. decorative, illustrative, navigational)
- Image Intent refers to the communicative goal of the image (e.g. supporting content, guiding interaction, evoking emotion).

It is argued that by defining alt text context in this structured manner, a more precise and meaningful approach to generating alt text is ensured. This definition moves beyond the vague notion of "context an image is used in" and provides a systematic way to assess when alt text is necessary and what it should convey. In particular, it is framed within two important elements

related to the image (type, function and intent) (Desai *et al.*, 2021; W. Chen *et al.*, 2023), and to the webpage (topic and purpose) (Mangiatordi and Lazzari, 2018). Given the complexity and multidimensional nature of context, it is therefore proposed that this structured definition should serve as a foundational framework for future efforts in alt text generation, and it is also the definition that has been used in this work (see section 9.2.1.3). For completeness, and to help better clarify the use of altC prior to its use in this work, two worked examples are included below:

Example 1: A product image on an e-commerce website *(Website topic)*. This photograph *(Image type)* appears on a product-selling page *(Webpage purpose)*. It has no interactive function, as it not a link/button *(Image function)*, and it illustrates the product *(Image intent)*.

Example 2: A decorative image *(Image function)* in a travel blog *(Webpage topic)*. This photograph *(Image type)* appears on an the page for entertainment *(Webpage purpose)* and aims to complement the adjacent content *(Image intent).*

4.3.2 Summary

Alt text plays a crucial role in web accessibility, traditionally catering to blind users via screen readers. However, this section discussed that people with diverse impairments also rely on alt text, necessitating a broader, more inclusive approach. Despite numerous efforts to define and improve alt text suitability, challenges persist, including misconceptions equating alt text to captions, context-dependent variations, and a lack of standardisation in automatically generated descriptions. Existing frameworks stress the need for structured, context-driven annotations, yet current practices fail to fully address contextual complexities. To bridge this gap, in this section, a structured semantic definition of Alt Text Context (altC) is proposed, which provides a systematic approach to evaluating and generating suitable alt text. Given the importance of structured annotation in achieving high-quality alt text, the next step is to examine different approaches to alt text annotation and evaluation.

4.4 Related work on alt text annotation and evaluation

This section shifts focus to alt text annotation and evaluation approaches, exploring how the aforementioned elements of context, as well as the concept of suitability are considered in state-of-the-art approaches.

4.4.1 Automated approaches

The need to scale alt text annotation in tandem with AI-related advancements resulted in automated annotation approaches, which have used several techniques to address missing and unsuitable alt text whilst scaling the task. However, despite their promise, past work also highlighted suitability and privacy concerns (Wu *et al.*, 2017), with the latter being most evident in social media contexts where people are depicted (Hanley *et al.*, 2021). Indicatively, *Microsoft* and *Google* have employed automated systems to generate alt text and suggest that in certain cases AI technology has propelled past the need to rely on human judgment for alt text suitability (Mazzoni, 2023; Roach, 2020). Additionally, Bennett *et al.* (2021) support that privacy concerns are not restricted to automated approaches. For alt text evaluation, automated evaluation tools are used, with their greatest aptitude being the speed at which they can check web content against the WCAG to identify alt text barriers, among other barriers (Kaur and Kumar, 2015a). Nevertheless, and as detailed in the previous chapter, conformance to the WCAG is far from a complete picture of accessibility, not least in relation to suitability of alt text. Evaluation tools have further been shown to lack transparency with regard to how they operate to detect barriers (Petrie and Bevan, 2009; Moreno *et al.*, 2011). Taken together, these bode well for proposing the CxDAI framework in the previous chapter to advance such evaluation tools, allowing for increased user agency. It is thus evident that automating alt text annotation and evaluation entails scalability benefits; however, those come at the expense of suitability and privacy, which grow prohibitive as the approach scales.

Nonetheless, suitability remains a pending issue with automation, not least in relation to more complex depictions, such as infographics (Zong *et al.*, 2022). Importantly, Birhane, Prabhu and Kahembwe (2021)'s audit of the output of automated approaches showed that the further such approaches are scaled, the further the compromise in alt text suitability and privacy. It is vital, however, that AI models used to automatically generate alt text are trained on very large datasets to function properly (Sharma *et al.*, 2018). Recent work has proposed AI models known as Vision-to-Language (V2L) captioning models, which use encoder-decoder methods: an image is inputted to the encoder processing the visual information in the image and passing it to the autoregressive language decoder generating the caption (Ramos et al., 2023). These models are pre-trained on datasets of image-alt text pairs, with the alt text being traditionally authored by humans (Schuhmann et al., 2022).

However, the resulting datasets were too small in size, negatively impacting the performance of the models; hence, newer models capitalised on the availability of a wide range of images

on the Web, retrieving web images and their associated alt text descriptions via web scraping (Changpinyo et al., 2021). Alt text is particularly sought after for its shortness, making it cost-effective to retrieve and to capture a more accurate snapshot of the visual information in the image (Laurençon et al., 2023). However, and although studies exist in the literature that focus on the suitability of automated approaches via V2L captioning models and relevant datasets, there is a lack of context in training datasets, as well as a lack of evaluation of the performance of the models on accessibility (see section 6.3).

In effect, the performance of state-of-the-art (SoTA) V2L models in terms of accessibility is bound to the noisiness of web-scraped datasets used to train them. This is evident in models like PaLI-17B (X. Chen *et al.*, 2023), which is trained on WebLI (ten billion images and tens of billions of image-alt text pairs), achieving SoTA results on captioning benchmarks, but the automatically generated alt text is prohibitive for accessibility due to inheriting the noise and lack of context of the scraped source. Other V2L models like IDEFICS (Laurençon et al., 2023) were trained on datasets that sought to incorporate context through the extraction of the image's neighboring text content, and links embedded in the image. However, the dataset is similarly collected from large-scale raw web corpora, inheriting limitations, such as incorrect grammar, incomplete or misleading metadata, stereotypical notions, and lack of context in alt text. This is also the case in lightweight approaches like SMALLCAP (Ramos *et al.*, 2023), i.e., a model that mitigates the need for fine-tuning across domains, because it does so by using a retrieval-based system that draws from web-scraped datasets to access general knowledge that it can then adapt according to different domains. It is therefore imperative to revisit the quality of the training data, which remains the key limitation in the performance of SoTA V2L models in terms of accessibility, as although automation is vital to scale the generation of alt text, it cannot address its suitability. In response, in this work, the need to automate alt text generation through AI models to address the reluctance of authors (Williams *et al.*, 2022) and scale the approach is acknowledged, but a community-driven, human-centred approach is used to gather training data (see section 10.2) to address the above suitability- and context-related limitations.

4.4.2 Manual approaches

The previous section showed how scaling the generation of alt text can result in non-negligible compromises in the context of suitability. Manual evaluation is, therefore, most often used to refine automatically generated alt text; nevertheless, Mack *et al.* (2021) have suggested that authors write more suitable alt text when they do so from scratch, but this is not always the case for more complex depictions, such as graphs and charts (Gleason, Carrington, *et al.*, 2019;

Chintalapati, Bragg and Wang, 2022). On complex depictions, Singh *et al.* (2024) used AI to help authors write better alt text for figures in academic publications by providing them contextual information, such as figure type and caption, mentioning paragraphs, extracted text, and data tables. Williams *et al.* (2022) also explored alt text for figures in academic publications, noting how suitability of such alt text varies per contextual information, such as the domain, the presence/absence of a caption and the figure type, with the latter ranging from diagrams, tables and text blocks, data visualisations, and other images. Thus, manual approaches are considered when suitability is paramount; however, Das *et al.* (2024) compared how 16 web content creators and 16 screen reader users evaluated and authored their own alt text for AI-generated images, highlighting a mismatch between their views on suitability.

In past work where manual evaluation is used to refine the automatically generated alt text, it has been shown that experts' views on the suitability of alt text can vary substantially (Mack *et al.*, 2021). This disparity in accessibility experts' views on alt text suitability is in fact not new (Lengua, Rubano and Vitali, 2022), and bodes well for proposing the altC definition in the previous section as a common blueprint on which to draw from. Past work has further shown that the recruitment of accessibility experts is not always affordable (Abuaddous, Jali and Basir, 2016), and given the aforementioned disparity in expert judgements on suitability, outsourcing alt text generation to larger groups of non-experts has been proposed as a solution.

4.4.3 Crowdsourcing approaches

The appreciable suitability bottlenecks mentioned in previous sections coupled with the need to scale alt text generation have led to a shift to crowdsourcing approaches, where annotation is handed out to large non-expert crowds in an effort to approximate the suitability of the manual approach whilst at the same time scaling annotation. However, the required resources must not be underestimated, as previous work highlighted the need for five non-experts to boast suitability at the level of one web accessibility expert (Vigo, Brown and Conway, 2013). On that note, the approaches that have been proposed in the literature span from friendsourcing (Brady *et al.*, 2013) to microworking incl. monetary incentivisation (Salisbury, Kamar and Morris, 2017) and have shown promise for suitability, but ironically, concerns were raised over whether they can work at scale (Brady, Morris and Bigham, 2014; Gleason *et al.*, 2020). Morash *et al.* (2015), on the other hand, showed how structured queries and templates can help improve the suitability of alt text written by crowdworkers for figures in academic publications by providing contextual information, such as figure type, table data and colour information. However, these approaches predate the recent surge in large-scale AI models (Mitchell, 2021),

and are largely in deficit (Bellscheidt *et al.*, 2023), despite recent research (Stangl, Morris and Gurari, 2020) suggesting the provision of context-related guidance to crowdworkers to train AI models with implications for suitability at scale. Indeed, recent efforts (e.g., Gleason, Pavel, *et al.*, 2019; Gleason *et al.*, 2020) reveal that, given the complexity of authoring suitable alt text, training of non-expert crowdworkers can only be foregone when authoring alt text for memes, which are considerably less hard to describe suitably.

This agrees well with past crowdsourcing work suggesting that training crowdworkers is essential for quality output for more difficult tasks (Madge, 2020), and it is in tandem with the aforementioned need for context-driven training prior to alt text annotation and evaluation (see section 4.3.1); hence, the proposed altC definition to help guide such training. However, current approaches, namely microworking and friendsourcing, appear lacking in incorporating training, owing to shortcomings in participation motivators (i.e., minor remuneration in microworking, social media contact support in friendsourcing) (Droutsas, 2021). Past crowdsourcing work for tasks that are considered similarly difficult to alt text annotation and evaluation (e.g., Yu *et al.*, 2022; Lafourcade and Le Brun, 2023) suggests *Games-With-A-Purpose* (GWAPs), i.e., the game-based crowdsourcing approach, to scale annotation and incorporate training, owing to the motivator for participation being gameplay enjoyability. However, and although GWAPs have been previously used for accessibility (e.g., Von Ahn *et al.*, 2006, 2007), there is no GWAP for alt text annotation or evaluation, let alone the incorporation of context-driven training of non-expert crowdworkers for suitability.

4.4.4 Summary

Alt text annotation approaches face a persistent trade-off between scalability and suitability. Past work showed that automated methods offer efficiency but often compromise contextual accuracy and pose ethical considerations. Manual annotation, on the other hand, has shown to improve quality but it lacks scalability, while crowdsourcing methods, even though they were a promising approach to mitigate this issue, they still require significant resources due to the high skill ceiling of alt text annotation. It is therefore necessary to train non-expert annotators in authoring alt text suitably, but existing approaches are ill-advised to support training. Past work suggests crowdsourcing games, i.e., GWAPs, as a promising approach for more complex annotation tasks; however, there are no GWAPs for alt text annotation in the literature.

4.5 Chapter summary

In this chapter, the literature landscape on alt text, its suitability, and the importance of context

for suitability were explored. The discussion revealed a mismatch between the views of web content creators and screen reader users, as well as a non-trivial disparity between what each accessibility expert considers suitable alt text. Further to this mismatch, and despite the general consensus on the need for structured, context-driven annotations, context in the case of alt text is loosely defined in current practices. In response, a structured semantic definition of Alt Text Context (altC) was proposed (see section 4.3.1) to help unify views and practice on when alt text is needed and what it should convey. The chapter further examined different approaches to alt text annotation and evaluation, pinpointing scalability and suitability deficiencies in automated and manual approaches, respectively. Taken together with the variability in experts' opinions on what makes alt text suitable, crowdsourcing approaches that hand out alt text annotation to large crowds of non-experts were proposed. However, crowdsourcing work on alt text is scarce, and the ability of such approaches to train non-expert crowdworkers is necessitated, as alt text annotation is prohibitively difficult without prior expertise. Whilst suitability of alt text cannot be compromised and expertise on suitability varies, it is vital that crowdsourcing approaches that afford training of non-experts in authoring alt text suitably based on a common blueprint, such as the proposed altC definition, are explored.

Nonetheless, a further challenge is the reluctance to author alt text (Williams *et al.*, 2022), and the need to scale alt text generation on par with advances in technology (Birhane, Prabhu and Kahembwe, 2021), which necessitate automating the process. GWAPs can, therefore, be used to gather a human-curated dataset via crowdsourcing, which can then be used to train AI models that automatically generate alt text. The complexity of authoring alt text suitably and the need to scale annotation favour GWAPs over other crowdsourcing approaches, as they can train users and motivate participation via gameplay enjoyment rather than monetary incentives, respectively (Tuite, 2014; Kicikoglu *et al.*, 2020). Accordingly, Fig. 4 below highlights key challenges discussed in this chapter in the case of alt text and current solutions, mapping these challenges to the approaches discussed in the following chapters.

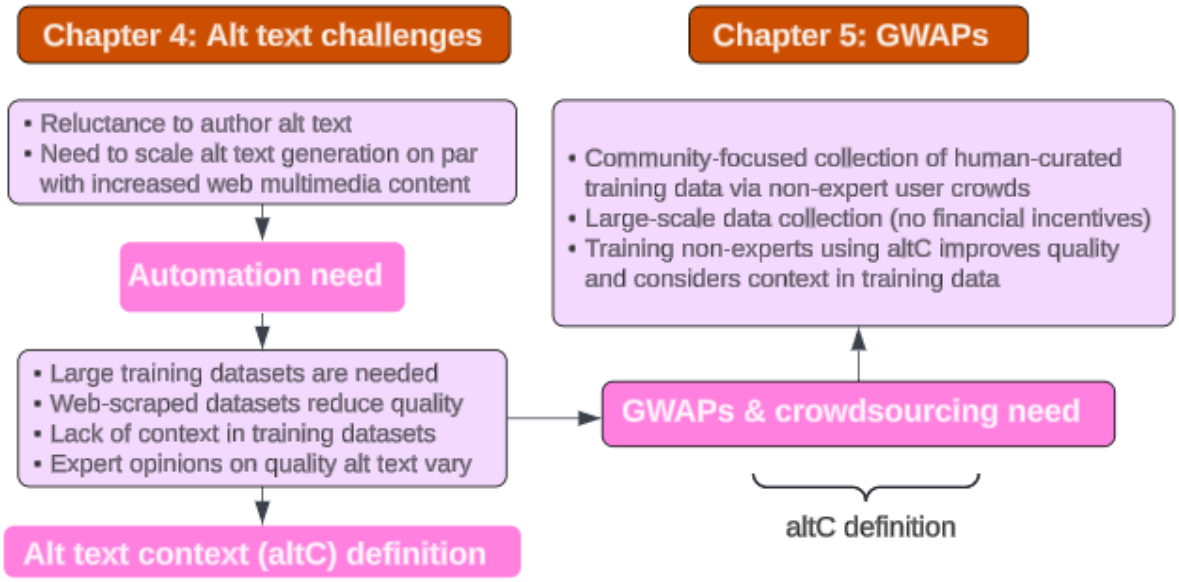


Fig. 4. Diagram mapping discussed alt text challenges (Chapter 4) to discussed solutions (Chapter 5)

Deferring to the above figure, it is vital that the generation of alt text is automated, as there is a reluctance to manually author alt text and a need for scalability. However, SoTA V2L models compromise quality whilst scaling alt text generation and large datasets are needed to train such models. Crowdsourcing is, thus, proposed to task non-experts to author alt text to gather such datasets, which also addresses the reported lack of consensus between accessibility experts on what constitutes suitable alt text. GWAPs are proposed as the most promising crowdsourcing approach to achieve this, as they do not motivate participation via financial incentives and excel at training players. The lack of primary financial incentives is a further strength of GWAPs to allow for large-scale data collection (Chamberlain *et al.*, 2013), which is vital for V2L models that require very large training datasets. To address the lack of context in existing approaches, the definition of alt text Context (altC) proposed in this chapter is suggested as a valuable blueprint for training non-experts in a GWAP context. Then, the GWAP output can be used to train V2L models, whilst aiming to approximate human-level quality and preserve context in automatically generated alt text (see chapter 11). Accordingly, Chapter 5 discusses GWAPs and their potential to train users and scale annotation.

Chapter 5. Games-With-A-Purpose (GWAPs)

5.1 Introduction

The previous chapter stressed the need to generate alt text at scale without compromising on its suitability, with crowdsourcing solutions, particularly Games-With-A-Purpose (GWAPs), being suggested to train crowdworkers, owing to previous success for more complex tasks (Lafourcade, 2020). Crowdsourcing was defined by Howe (2008, p. 1) as "the act of taking a task traditionally performed by a designated agent (such as an employee or a contractor) and outsourcing it by making an open call to an undefined but large group of people." It is rooted in Surowiecki (2005)'s *wisdom of crowds* concept, which advocates that given a large number of non-experts, collective decision-making can outpace subject matter experts. Crowdsourcing also maps well to the need to scale alt text annotation by handing it to large non-expert crowds, which are preferred over expert opinions in the case of alt text, where they have been shown to vary substantially per individual expert (Lengua, Rubano and Vitali, 2022). The approach also aims to not compromise on quality, as alt text is still being authored by humans. However, it has been shown that for more complex tasks, as is the case with authoring alt text suitably, handing the task over to non-experts without prior training is overkill. This is addressed in this work by using the proposed definition of context in alt text (see section 4.3.1) to guide the tutorial phase of the implemented GWAP solution (see section 9.2.1). Indicatively, elements from the altC definition are mapped to values of in-game context prompts (Table 10), and are then assembled into a **natural language prompt** in the game based on a structured template, which is discussed in detail in Section 9.2.1.3.

In effect, crowdsourcing tasks have grown increasingly ambitious with the passage of time; thus, the ability of the approaches to incorporate training of non-experts is vital. Further, such approaches are distinguished from each other on how they motivate participation, which also determines their scalability, as crowdsourcing approaches rely on their ability to recruit large crowds to scale annotation. The few crowdsourcing approaches that have been proposed for alt text annotation and evaluation (see section 4.4.3), however, did not show promise with regard to trainability and scalability, owing to easier tasks (i.e., authoring alt text for memes on Twitter) and non-scalable crowdsourcing variants (i.e., friendsourcing). Past literature on complex tasks and crowdsourcing, ranging from linguistic annotation to protein folding (Curtis, 2015; Madge *et al.*, 2022), highlights GWAPs as the most promising such approach for training non-experts and scaling annotation. However, there are, to the best of the researcher's knowledge, no GWAPs for suitable alt text annotation and evaluation. Accordingly, this chapter presents a discussion on crowdsourcing and intertwined concepts, as well as an

overview of GWAPs as an approach to more complex crowdsourcing tasks, followed by an appreciation of such state-of-the-art GWAPs for alt text annotation that is both suitable and scalable.

5.2 Crowdsourcing, intertwined concepts and variants

Crowdsourcing is tied to the concept of *Collective Intelligence* (CI), i.e., "*groups of individuals doing things collectively that seem intelligent*" (Malone, Laubacher and Dellarocas, 2009, p. 2). Recent research has, in fact, shown that CI emerges from collaboration and is only evident at an aggregate level (Dortheimer, 2022; Kameda, Toyokawa and Tindale, 2022); hence, its tie to crowdsourcing. It has further been suggested that CI emerges most evidently in crowds with diverse expertise on a subject matter, as it fosters creative decision-making (Dellermann *et al.*, 2020). It can thus be surmised that crowdsourcing is a possibility space leveraging Surowiecki (2005)'s aforementioned *wisdom of crowds* to allow for CI to emerge. However, reliance on crowd-based decision-making over subject matter experts can be problematic (Chafetz, 2005), as individual decisions within heterogeneous crowds can be more noise than signal; thus, it is important to distinguish between wise and unwise crowds. This is addressed in this work by using a common blueprint, namely the proposed context definition (Section 4.3.1), to train non-expert alt text authors (unwise crowd) towards becoming pseudo-experts (wiser crowd).

In this vein, Surowiecki (2005) outlined four criteria, i.e., diversity of opinion, decentralised decision-making, member independency and judgment aggregation, for distinguishing between wise and unwise crowds. Of these, diversity of opinion and member independency map most well to alt text annotation, where non-biased plural opinions are preferred over expert opinions (Cooper, 2022), due to the latter's substantial variability (see chapter 4). Alt text annotation is, in this regard, a human computation task; i.e., short tasks where expertise on a subject matter is not expected and are thus easy for humans but hard to automate for computers (Quinn and Bederson, 2011). Handing alt text annotation over to non-experts, however, is challenging for suitability (see section 4.2), as although the task is easier for humans than for computers, it has been proven complex even for accessibility experts. Given the reported discrepancy in experts' takes on alt text suitability, however, it is important to gather large non-expert crowds; and, in this work, a redundancy mechanism, namely assigning the same task to many non-expert alt text authors to capture diverse opinions, is used to address the complexity of alt text suitability (see section 9.2.1.1).

Whilst it is evident that trainability is required in crowdsourcing solutions to alt text barriers, it is important to note that the convergence of human computation and crowdsourcing was first

motivated by the need to scale the former with the growth of the Web (Howe, 2008). As shown in the previous section, the scalability of crowdsourcing approaches is linked to their incentives for participation, which Malone, Laubacher and Dellarocas (2009) categorised into variations of *money, love* and *glory*, referring to reasons why people participate in approaches where CI may emerge. The core incentive in microworking, which has been used for alt text annotation (see section 4.4.3), is *money*; i.e., "The promise of financial gain … Sometimes people receive direct payments" (Malone, Laubacher and Dellarocas, 2009, p. 5), ranging around 1.65 to 2.53 pounds per hour (Molina *et al.*, 2023), which is prohibitive owing to the proportionate cost increase of the approach as annotation scales. Further concerns have, in fact, been raised with microworking, such as a scholarly divide on its moral dimensions and its social insecurity, not least in relation to the vulnerability of microworkers when they want to take a break, leave or retire (Klaus, Haas and Lamura, 2023).

Friendsourcing, which has also been used for alt text annotation (see section 4.4.3), motivates participation through *love*; i.e., "the opportunities it provides to socialize with others" (Malone, Laubacher and Dellarocas, 2009, p. 5), as it narrows down outsourcing of annotation tasks to close communities with access to otherwise unattainable information (Rubya, Numainville and Yarosh, 2021). Whereas scalability is, therefore, severely compromised with the friendsourcing approach, it has been shown that the increase in quality is negligible compared to microworking for certain tasks (Bateman *et al.*, 2017). Citizen Science, namely non-scientists' participation in projects initiated by scientists, is a crowdsourcing approach that, although it has yet to be used in the case of alt text, has shown promise for large-scale annotation (Hecker *et al.*, 2018). It incentivises participation through *love*; i.e., "it makes them feel they are contributing to a cause larger than themselves" (Malone, Laubacher and Dellarocas, 2009, p. 2), but it does not map well to alt text annotation, as it aims to recruit subject matter experts instead of training non-experts. On the other end, GWAPs use *love*; i.e., "people can be motivated by their *intrinsic* enjoyment of an activity" to motivate participation (Malone, Laubacher and Dellarocas, 2009, p. 5), and although scholars agree that they have shown promise for more complex and large-scale tasks, they have, to the best of the researcher's knowledge, yet to be used for alt text annotation and evaluation.

5.3 Games-With-A-Purpose (GWAPs)

GWAPs are the game-based crowdsourcing approach pioneered and defined with the release of the *ESP* game for image annotation (Von Ahn and Dabbish, 2004), as computer games that people play and "could, without consciously doing so, simultaneously solve large-scale

problems" (Von Ahn, 2006, p. 92). As discussed in the previous section, this approach appears most promising to scale annotation, as it motivates participation via intrinsic enjoyment of gameplay, which is cost-free and more ethically sound than financial compensation. GWAPs are, in this vein, less inviting of malicious behaviour due to the absence of financial incentives (Lyding, Nicolas and König, 2022), and they also excel at structuring training ((Tuite, 2014); thus, they map more naturally to more complex annotation tasks. Although they have yet to be used for alt text annotation, it has been shown that GWAPs do outperform other crowdsourcing approaches as task complexity increases, such as with linguistic annotation tasks (Madge *et al.*, 2022). Increased task complexity, however, overlaps with the flexibility to allocate design choices towards gameplay enjoyment (Kicikoglu *et al.*, 2019), compromising GWAPs' main participation incentive and selling point. It is therefore vital that taxonomies and frameworks for the design of GWAPs are explored to address such limitations.

5.3.1 Typologies and design frameworks

By design, GWAP gameplay is infinite, revolving around a core loop (Games and NLP, 2019), which players need to find enjoyable to continue performing annotations and thus contribute to the game's intrinsic purpose. This has led scholars (Krause and Smeddinck, 2011; Madge, 2020) to assimilate GWAPs with incremental games, such as *Free-to-Play* (F2P) games, *'Ville games* and *social network games* (SNGs), that also revolve around an infinite core loop. In fact, Madge (2020) proposed a typology of such games highlighting their shared characteristics, such as requiring no initial commitment or payment and being designed to achieve returns via ongoing player contributions, as well as their need to attract and retain large player bases for scalability. A fundamental distinction between GWAPs and these games, however, is that the former need to achieve returns, i.e., "solve large-scale problems", vicariously, i.e., "without consciously doing so", as part of gameplay enjoyment (Von Ahn, 2006, p. 92). Despite the enjoyability of GWAPs being, therefore, essentialised, it is neglected by current typologies, which focus more broadly on human computation (Quinn and Bederson, 2011).

On the other hand, Pe-Than, Goh and Lee (2015) proposed a more holistic typology of the design of GWAPs featuring three key dimensions, namely *gameplay mode, gameplay structure* and *data*. Nevertheless, it has been argued that more recent GWAP designs seldom draw on typologies or design frameworks; instead, they tend to incorporate simple strategies that had shown promise in previous GWAP designs (Siu, Zook and Riedl, 2017). This is, for the most part, due to the success of the aforementioned ESP game for image labelling, resulting in its redevelopment by Google into the Google Image Labeller (Chamberlain *et al.*, 2013), which

remains unmatched by current GWAPs both in terms of gathered annotations and recruited players (Droutsas, 2021). In this vein, the designers of the ESP game proposed three strategies for GWAP designs for gameplay enjoyment and quality output; i.e., input-agreement, output-agreement and inversion-problem strategies (Von Ahn and Dabbish, 2008). The output-agreement strategy, which was also used in the ESP game, tasks two players with making the same annotation judgement independently; the game will only consider their judgements as correct if they are consensual.

Contrastingly, GWAPs that use the input-agreement strategy (e.g., Law *et al.*, 2007; Chrons and Sundell, 2011) present players with inputs known to be identical or unidentical, awaiting for their consensus on whether they had been given identical inputs or not. Finally, GWAPs that use the inversion-problem strategy (e.g., Parasca *et al.*, 2016) assign roles to players; i.e., one player is the *narrator* and is tasked with making an annotation judgement, and then to clue in *guessers*; i.e., players tasked with matching the former's judgement. As discussed in the previous section, however, increased task complexity is a further challenge in designing for gameplay enjoyment, as opposed to, for example, image labelling in the ESP game, which is easy for humans without prior expertise. On more complex tasks, *ZombiLingo*[4] is a GWAP for dependency syntax annotation, which draws on an adaptation of the MICE (Money, Ideology, Constraint and Ego) framework (Burkett, 2013); i.e., money is replaced by reward, ideology by interest, constraint by light constraint and ego by player role in relation to the rest of the playerbase (Fort, Guillaume and Chastant, 2014). Although ZombiLingo has been successful in gathering thousands of annotations for such a complex task, its design has been critiqued as too similar to typical GWAP designs like the aforementioned strategies (Madge, 2020).

Further on complex tasks, *PhraseDetectives* for anaphora resolution and *JeuxDeMots* for lexical relation extraction are some of the most successful GWAPs, having gathered millions of annotations (Yu *et al.*, 2022; Lafourcade and Le Brun, 2023). Other GWAP designs have used player progression with varying evidence on its effectiveness (Madge *et al.*, 2019, 2022; Kicikoglu *et al.*, 2020). More relevant to this work, Bellscheidt *et al.* (2023) proposed a set of design recommendations for GWAPs for alt text annotation; however, **none** of the GWAPs that the work draws on were designed for alt text annotation. In response, the GWAP solution developed and implemented in this work draws on Pe-Than, Goh and Lee (2015)'s typology of GWAPs, as well as on Bellscheidt *et al.* (2023)'s design framework for alt text annotation in a GWAP context (see section 9.2.1.1). The former is chosen for its focus on games' design,

[4] https://www.zombilingo.org

including gameplay mode, structure and data (see Table 9) rather than GWAP purpose, while Bellscheidt *et al.* (2023)'s framework is chosen for being the only design framework tailored to alt text annotation in a GWAP context.

5.3.2 Metrics

Whilst it is evident that GWAP success is measured in terms of annotation quantity and quality, and gameplay enjoyability, Von Ahn and Dabbish (2008) proposed the initial relevant metrics:

> *Throughput*: Average completion rate of annotations per human hour
> *Average lifetime play (ALP)*: Average overall gameplay time by each individual player
> *Expected contribution*: Throughput * ALP

The authors recognised the lack of measures for enjoyability in these metrics and Siu, Zook and Riedl (2017) later divided GWAP success metrics into *player experience* and *task completion* metrics. More recently, Madge (2020) expanded on the first metrics (Von Ahn and Dabbish, 2008), proposing a contemporary adaptation; indicatively:

- *Cost per Judgement* (CpJ): Average cost for a meaningful annotation judgement.
- *Cost per Item* (CpI): Average cost for a completely annotated item. This can vary per GWAP and is, naturally, requiring multiple annotation judgements, and it can depend on how well task to user assignment, player trainability and task presentation is done.
- *Lifetime Judgements* (LTJ): The total number of overall judgements per individual player divided by the total number of players.

5.3.3 GWAPs for alt text annotation

As previously discussed, the unmatched success of the ESP game led to newer GWAPs drawing on its design instead of typologies or design frameworks, which is also the case for GWAPs aimed at accessibility. More relevant to this work, *Phetch* was a GWAP for alt text annotation using images labelled by the players of the ESP game, but its in-game task resembled annotation for image captions rather than alt text (Von Ahn *et al.*, 2006, 2007). However, their reported benefits inspired further similar work (Yuan, Sapre and Folmer, 2010; Steinmayr *et al.*, 2011; Nguyen *et al.*, 2019, 2020; Harris, 2020), but it must be noted that their utility was identified in sourcing crowds to effectively annotate content, which again resembled keywords rather than alt text. In terms of suitability and context of alt text, Mangiatordi and Lazzari (2018) proposed a gamified plugin for crowdsourced annotation that allows for alt text

annotation that is aware of the context in which each image is used in; however, it must be noted that the authors would need to manually add contextual information to a database, which the plugin would then use to improve the suitability of the alt text. Although their approach essentialises the concept of 'context in which the image is used in' (see section 4.3.1), it fails again to move beyond this generic notion of context. To the best of the researcher's knowledge, there are no additional relevant studies, further highlighting a literature gap for similar solutions.

5.4 Chapter summary

This chapter drew on the call for a GWAP approach to gather a dataset for training an AI model that automatically generates alt text descriptions, surveying crowdsourcing literature to better understand crowdsourcing as a concept. This allowed for a comparison of the adequacy of GWAPs for alt text annotation and evaluation compared to other crowdsourcing approaches in terms of trainability and scalability. The interconnection of crowdsourcing with CI and human computation was thus discussed, revealing that the scalability of crowdsourcing approaches is linked to their core incentives for participation. Accordingly, four such approaches were discussed for addressing alt text barriers (see Table 2 below).

Table 2. Crowdsourcing approaches per incentive and scalability potential

Approach	Core incentive	Annotation scalability	Benefits	Drawbacks
Citizen science	Scientific contribution	Large scale	Cost-effective, democratisation	Reliant on recruiting experts and cannot afford training of non-experts
Microworking	Monetary compensation	Small to medium scale	Time-effective, mix of job and joy	Prone to ethical issues and costly for large-scale annotation
Friendsourcing	Social contribution	Very small scale	Annotation nuance	Limited to personal social media network and prone to social costs
GWAPs	Gameplay enjoyment	Large scale	Cost-effective, training non-experts	Reliant on sticking to gameplay enjoyment to motivate participation

Deferring to the above table, GWAPs and citizen science appear most promising to address the need to scale annotation; however, the latter does not map well to the need to train non-experts, which is essential due to the variability of expert views on alt text suitability. It was further discussed that although GWAPs have shown promise for more complex tasks, as well as for tasks similar to alt text annotation, there is a gap in GWAPs for alt text annotation in relevant literature. Whereas a plurality of GWAP design frameworks and typologies were discussed in

this chapter, it was shown that these are often neglected by newer GWAPs, which draw on previous successful designs instead. On the other hand, in this work, Pe-Than, Goh and Lee (2015)'s typology, and Bellscheidt *et al.* (2023)'s design framework are used for the proposed GWAP solution (see section 9.2.1.1). Indicatively, gameplay mode is asynchronous, as players author alt text and rate other players' alt text independently, with the game's structure being that of a standalone single-player game, while players contribute collaboratively through data annotation. The approach to data annotation and evaluation is open-ended and aimed at data redundancy through the collection of a plurality of alt text and rating scores from different players. A detailed mapping of the GWAP proposed in this study and its related mechanics with Pe-Than, Goh and Lee (2015)'s framework can be seen in Table 9.

It is, however, acknowledged that this structure may introduce some biases that need to be taken into account. First, gameplay being asynchronous can potentially lead to rating score variance across different players that relates to misunderstandings, rather than explicit alt text quality differences. This is particularly the case considering that GWAP players are not experts in alt text annotation before interacting with the game, with the goal being for the game to train them into pseudo-experts. The use of context prompts can bias players towards focusing on specific aspects instead of freely describing the images based on their understanding of suitable alt text. This is despite providing a common blueprint under which players are trained, as there are no reported measures of how players could interpret the components of context prompts. Nevertheless, this open-ended, asynchronous structure captures diverse opinions and addresses the variability in expert judgements on alt text suitability (Hanley *et al.*, 2021).

However, the promise of GWAPs for suitability stems from the trainability of the approach, and in the case of alt text, it is imperative that training draws on a common blueprint, such as the proposed Alt Text Context (altC) definition (see section 4.3.1), to mitigate discrepancies in suitability views. Therefore, GWAPs must evolve to incorporate deeper contextual awareness for suitability but also ensure that annotations are generated at a sufficient scale for training AI models. As discussed in the previous chapter, it is important to automate alt text generation through V2L captioning models, which need very large training datasets to function properly. Whilst the key limitation in the performance of the models is the noisy training datasets that are typically web-scraped, GWAPs were proposed as a community-driven approach that is human-centred as an alternative for large-scale training data collection. However, to motivate the GWAP approach for the training of V2L models, it is first necessary to appreciate SoTA models and the datasets they are trained on. Accordingly, Chapter 6 synthesises available literature on this topic.

Chapter 6. Vision-to-language (V2L) models

6.1 Introduction

The previous chapters highlighted the need to scale alt text annotation and evaluation while considering the reluctance of web content creators to author alt text (Mack *et al.*, 2021), as well as the difficulty of the task (Hanley *et al.*, 2021). The recent surge in large-scale AI models has increased the demands on scalability; however, state-of-the-art (SoTA) automated approaches are prohibitive in terms of alt text quality (see section 4.4.1). On alt text quality, the concept of context in which the image is used in has been deemed essential; hence, a definition of this concept was proposed (see section 4.3.1). Whilst it was discussed in the previous chapter that V2L models need very large training datasets, GWAPs were proposed as a community-driven alternative to the web scraping paradigm to address large-scale training data collection. Indeed, the noisiness of web-scraped training data was highlighted as a key limitation of SoTA V2L models to generate alt text that is relevant for accessibility (see chapter 5). Although GWAPs have shown promise to scale annotation due to the lack of financial incentives and boast quality alt text due to being able to incorporate context-driven training, it is first important to explore current V2L models and the datasets they are trained on to identify gaps in SoTA automated generation of context-driven alt text. Accordingly, this chapter presents a discussion on current V2L models, which process images and translate visual information into text descriptions, with a focus on alt text descriptions and the incorporation of context.

6.2 V2L model datasets

As previously discussed, V2L models use encoder-decoder methods to generate captions and are trained on datasets of image-alt text pairs that used to be authored by humans, but the surge of AI models has led to a paradigm shift, with the now most common practice being the use of web-scraped Internet-sized datasets (see section 4.4.1). However, it has been noted that web-scraped alt text descriptions are of particularly poor quality, and the images are removed from the context in which they are used in during web scraping, resulting in context-poor captions generated by the models (W. Chen *et al.*, 2023). There is a scholarly divide on this issue, with arguments, such as ‘scale beats noise’, suggesting minimal curation and the retrieval of larger quantities of poor-quality alt text via web scraping, while few research efforts are aimed at improving the quality of training datasets. In this vein, Laurençon *et al.* (2023) scraped web images and their neighbouring text content instead of their alt text to create OBELICS, a dataset comprising 141 million web documents and 298 million unique images, highlighting the

content the image depicts, the **text content in which it appears,** and **contextual information** as aspects that need to be preserved in such datasets.

Whilst OBELICS is one of the few efforts reported in the literature for preserving alt text context in training datasets, it is based on data extracted from Common Crawl, which is a gigantic dump of web-scraped data (Luccioni and Viviano, 2021), and thus inherits issues, such as bias, hate speech, noise and racism, from Common Crawl. Relatedly, Chen *et al.* (2023) noted the lack of **subject awareness** in existing datasets, using image clusters with high visual similarity and automatically generated descriptions for the subject of each cluster to create the seed dataset, which comprises two million such pairs. Further, Desai *et al.* (2021) introduced RedCaps, i.e., a dataset of 12 million image-text pairs from 350 subreddits, arguing that Reddit is an effective alternative to complex data curation, as content on this platform is meant and moderated for human interaction, resulting in more **emotionally rich** descriptions. Taken together, these efforts highlight the importance of context and its lack thereof in the datasets used for training V2L models, not least in relation to its impact on the quality of the text descriptions generated by these models. Additionally, such datasets are continuously critiqued for raising a plurality of ethical concerns relating both to the imagery included and the resulting captions, as well as for their neglect of accessibility (Birhane, Prabhu and Kahembwe, 2021). There is also a lack of human-curated alt text datasets with the potential to scale on par with the needs for training V2L models (Nguyen *et al.*, 2020). **No such datasets** have in fact been created via a crowdsourcing approach to address the dual challenge of lacklustre scalability and alt text quality with the manual and web scraping approaches, respectively. This is an important distinction and strength of the solution proposed in this work, which is based on a community-focused approach that is also human-centred by recruiting and training non-expert crowds to author alt text through a crowdsourcing game. Therefore, the use of a GWAP to gather training data distinguishes this work from efforts focusing on web scraping, owing to the resulting datasets being prohibitive for accessibility.

6.3 V2L models for alt text generation

The discussion in the previous section highlighted limitations in the datasets that V2L models are trained on as a key issue with regard to the quality of the generated text descriptions. This is crucial in the case of alt text, as it has been shown that unsuitable alt text can be equally or even more problematic than missing alt text (see section 3.3); therefore, the ability of current models to scale alt text generation becomes less relevant with regard to accessibility. PaLI-17B, for example, suggested a jointly scaled multilingual model, achieving SoTA performance

on captioning benchmarks, but the generated captions are not relevant for accessibility due to the noise and lack of context in the training dataset, i.e., WebLI (tens of billions of image-alt text pairs) (X. Chen *et al.*, 2023). Another model that is prone to the lack of context in web-scraped training data is InternVL-Chat that paired a gigantic 6 billion parameter vision encoder with a large language model through a large language middleware and a progressive training strategy, achieving SoTA performance on most benchmarks (Chen *et al.*, 2024). The suitability of generated alt text is similarly compromised in SMALLCAP, i.e., a lightweight model using a datastore to retrieve image-alt text pairs to lift the need for fine-tuning across domains, as the poor quality of alt text retrieved from the datastore is reflected in automatically generated alt text (Ramos *et al.*, 2023). Further, IDEFICS is a V2L model that learns from context-rich data, namely from the OBELICS dataset discussed in the previous section, showing an increase in performance via a simplified architecture and the processing of images at their native resolution and aspect ratios (Laurençon *et al.*, 2023). However, IDEFICS is also limited in its ability to automatically generate alt text that is relevant for accessibility, as it is trained on data extracted from the web-scraped Common Crawl dump.

Table 3 below includes SoTA models, the datasets they were trained on, and whether these datasets incorporated context, as well as their performance on common benchmarks. There are more AI models trained on these datasets; however, those were not included, as they were not fine-tuned for V2L captioning and were thus unrelated to the focus of this work.

Table 3. State-of-the-art AI solutions for alt text generation (alphabetically ordered)

AI Solution	Dataset	Context	Performance
IDEFICS (Laurençon *et al.*, 2023)	OBELICS	Yes	CIDEr COCO ZS: 91.8 VQAv2 ZS acc.: 60.0%
InternVL-Chat (Chen *et al.*, 2024)	6.03B image-text pairs	No	CIDEr COCO ZS: 142.4 VQAv2 ZS acc.: 71.7%
PaLI-17B (X. Chen *et al.*, 2023)	WebLI	No	CIDEr COCO ZS: 149.1 VQAv2 ZS acc.: 84.3%
SMALLCAP (Ramos *et al.*, 2023)	COCO (Lin *et al.*, 2014)	No	CIDEr COCO ZS: 119.7

The above table shows the lack of context in which images are used in the datasets that recent V2L captioning models are trained on. Notably, datasets that were discussed in the previous section for incorporating context, such as RedCaps and the seed dataset, have only been used for other applications, i.e., visual-text representation learning and image-text retrieval. The performance of V2L captioning models is evaluated on benchmark datasets, such as MS COCO (Microsoft Common Objects in COntext) and VQAv2 (Visual Question Answering), and reported via metrics like CIDEr (Consensus-based Image Description Evaluation). SoTA performances on these benchmarks are observed in PaLI-17B and InternVL-Chat, which are valuable from a computer vision standpoint; however, these are not relevant to accessibility, as the quality of the generated text descriptions is not evaluated for access via screen readers.

From a human-computer interaction (HCI) standpoint, two key gaps are identified in the automated generation of alt text descriptions via AI models, i.e., the lack of context in training datasets and the lack of evaluation of the performance of the models on accessibility. These gaps are addressed in this work by using context-driven training of non-expert alt text authors to gather training data (Section 9.2.1) and by evaluating the performance of models in terms of alt text quality and the ability to generate context-driven alt text (Chapter 11). The proposed models are therefore evaluated using the following metrics:

- *Human-perceived alt text quality* via player rating scores and statistical tests
- *Training effectiveness* via non-parametric inferential statistics and effect size estimation between trained and control versions of the models
- *Context presence* via binary presence evaluation of elements of the altC definition (see section 4.3.1) in automatically generated alt text

These metrics align with the need for evaluation to focus on accessibility and use participants' rating scores for human perceptions of alt text quality, agreeing well with recent accessibility studies (Kreiss *et al.*, 2022; Leotta, Mori and Ribaudo, 2023) on the importance of subjective evaluation of alt text. Training effectiveness and effect size estimation were measured via non-parametric statistical tests, as the data were not normally distributed (Section 11.2.3). Finally, context presence was assessed due to its discussed central role in the suitability of alt text (see chapter 4) and its lack thereof in alt text that has been automatically generated.

Importantly, it is clarified that the ML models used in this work (Chapter 9) were existing V2L models (T5-small and Qwen2.5-VL-3B-Instruct), which were fine-tuned, as opposed to architectural development. This choice was motivated by the aim of this work being to prove

a concept that is relevant to HCI and accessibility through the training of models, and it was thus deemed beneficial to resort to a more computationally efficient and easily available solution. Using existing models allowed for leveraging such models' pre-trained knowledge, i.e., encoding of visual understanding from corpora, such as COCO and Conceptual Captions, which could then be refined by the GWAP-generated data. Finally, and owing to the need for this work to contribute to HCI and accessibility, architectural novelty was not considered as a contribution; thus, focus was on the ability of the models to learn from human-curated and context-driven data to generate more suitable alt text for screen readers and their users.

However, it is recognised that despite the uniqueness of the GWAP dataset in this research with regard to crowdsourced data collection, training of non-experts and the use of context, the dataset introduces certain biases for model generalisability. Unlike web-scraped datasets, it is limited to the data collection that took place in this research and it can thus only function as a proof-of-concept. This is despite the data gathered not being as noisy as web-scraped datasets, due to being human-curated; however, participation being voluntary in this work (Chapter 10) can suggest interest in accessibility, which blurs the notion of a non-expert alt text author. Further, the evaluation of the performance of the models based on the accessibility-focused metrics proposed in this section was achieved through a further user study with former players of the GWAP (Chapter 11), but these players already had an understanding of alt text suitability owing to their previous gameplay experience. There is thus a trade-off between the scalability of the approach and the use of the GWAP dataset to train the ML models, as larger datasets are needed for the more reliable training of V2L models. Nevertheless, this work contributes two proof-of-concept models, demonstrating the value of context-aware, GWAP-generated human annotations to automate human-level alt text generation, which is unaddressed in prior work.

6.4 Chapter summary

This chapter discussed V2L captioning models and the datasets they are trained on, with a focus on their ability to scale alt text generation on par with modern needs while ensuring the quality of the generated alt text descriptions. It was shown that these models require very large datasets of image-alt text pairs for training to boast quality generated alt text. Although the scale of such datasets has increased, they seldom incorporate the context in which images are used, resulting in poor-quality alt text from an accessibility standpoint (see chapter 9). Additionally, and as discussed in the previous chapter, there is also a lack of crowdsourcing approaches that collect training data for V2L models. These approaches are, however, promising in principle to address both scalability limitations in the manual approach and quality limitations in the web scraping

approach. Crowdsourcing has in fact been used to refine the output of models trained on the RedCaps and the WIT (Wikipedia-based Image Text) datasets (Desai *et al.*, 2021; Srinivasan *et al.*, 2021); however, this has, to the best of the researcher's knowledge, yet to be attempted for the generation of alt text descriptions. A further revealed gap was the lack of evaluation of the performance of such models on accessibility (i.e., alt text description suitability); instead, evaluation is focused on computer-vision-related metrics. Fig. 5 below summarises the key challenges identified in the literature throughout this work, mapping them to the solutions that are proposed in this research to address them.

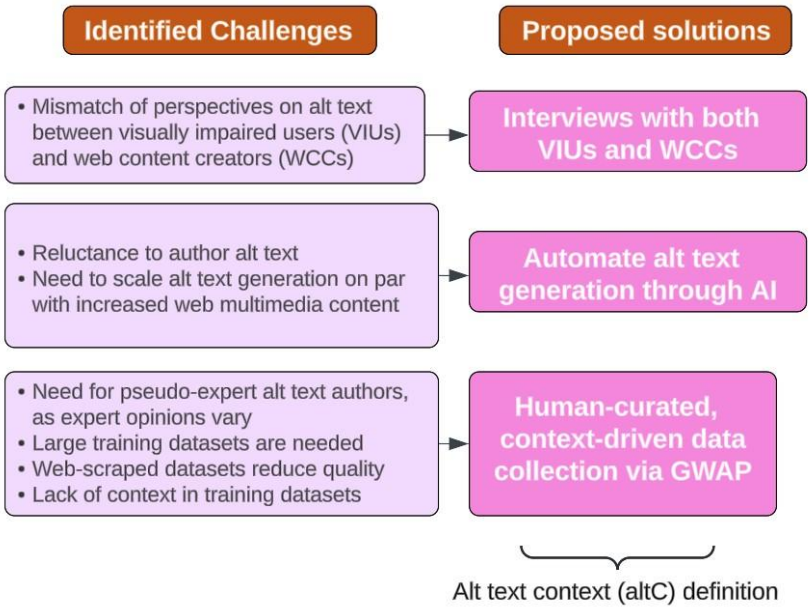


Fig. 5. Summary of key identified challenges in alt text research and proposed solutions

In response to the challenges shown in Fig. 5, a human-centred approach will be followed in this research, starting from the collection of qualitative user insights and their use to guide the design of a GWAP for context-driven alt text data collection, which will be used to train AI models to automate alt text generation. Whereas SoTA V2L models (e.g., InternVL-Chat and PaLI-17B) are evaluated on captioning benchmarks with no relevance for accessibility, the models proposed in this work are instead evaluated on the metrics discussed in Section 6.3. Accordingly, Chapter 7 discusses the methodological framework underpinning the research work in this thesis, whilst addressing the need to use a context-driven GWAP-based approach for gathering the dataset that will be used to train AI models for generating alt text. In response to the gaps of SoTA V2L captioning models identified in this chapter, the performance of models will be evaluated on the quality of the generated alt text, as well as on the representation of context in automatically generated alt text descriptions.

Chapter 7. Methodology

7.1 Introduction

The previous chapters presented a review of the web accessibility literature and the relevance of alt text barriers (Chapters 2 and 3), with existing efforts leaning towards crowdsourcing game-based solutions, i.e., GWAPs, to address such barriers, whilst it is vital to automate the approach (Chapter 4). The review of GWAPs in relation to other crowdsourcing approaches (Chapter 5) highlighted the adequacy of the GWAP approach to gather datasets for training AI models to generate alt text descriptions, as they are most apt for large-scale data collection and excel at training users. GWAPs are further proposed as a solution that is community-focused and human-centred to collect training data, as opposed to the web scraping paradigm, which is prohibitive for accessibility. This led to the exploration of SoTA V2L models and the datasets they are trained on (Chapter 6). Current models are trained on very large datasets deriving from web scraping; thus, alt text quality is in deficit, bearing little to no implications on improving web accessibility. It is therefore necessary to explore novel solutions for constructing alt text annotations that are suitable and at a speed and scale that is on par with modern accessibility needs. Accordingly, this chapter describes the methods and the theory that underpins them, that will be used in this work to address the discussed needs.

7.2 The research onion

The research onion concept (Saunders, Lewis and Thornhill, 2009) is used as a guiding diagram for the methods chosen in this research, split into six layers, that is, philosophy, approach to theory development, methodological choice, strategy(ies), time horizon, and techniques and procedures (from outermost to innermost) (see Fig. 6 below).

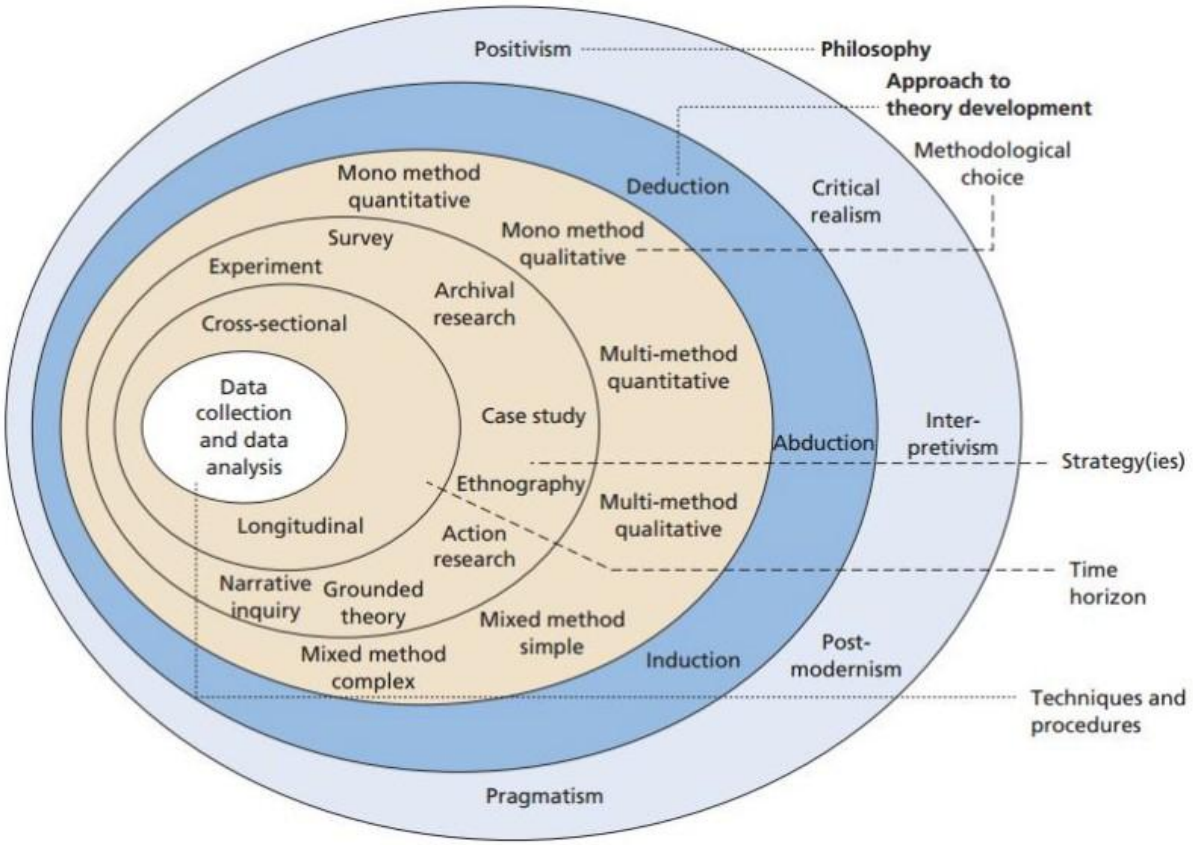


Fig. 6. The 'research onion' (Saunders, Lewis and Thornhill, 2009)

7.2.1 Outer layers

This section describes the aspects of the outer layers of the onion that are relevant to this work, underpinning the data collection and analysis methods chosen at the innermost layer, which are described in the next section.

7.2.1.1 Philosophy: Pragmatism

This research draws on pragmatism, a philosophy where the value of concepts is only deemed important when in tandem with action (Kelemen and Rumens, 2008). Pragmatist research starts with a research problem, which, in the case of this work, relates to the need for a novel solution to authoring and evaluating suitable alt text at scale, drawing on the deficiencies of the state of the art. Other philosophies were also considered, including positivism, interpretivism, critical realism and postmodernism. Positivism seeks objective measures that do not map well to the need to gather qualitative insights from those who create and those who consume alt text, while critical realism explores underlying causes that shape organisational structures and thus does not map well to the development of a practical solution to alt text barriers. Similarly, the need for a practical solution does not align with postmodernism, which aims to question and critique dominant ways of functioning within structures (Calás and Smircich, 2018). Interpretivism, on

the other hand, was not fully adopted again due to the need to tackle alt text barriers through a novel practical solution, but interpretivist elements were adopted. Indicatively, it was necessary to explore the subjective perspectives and experiences of those create and those who consume alt text, respectively (RQ2), owing to the mismatch between their views (Chapter 2). Contextualising this research within pragmatist philosophy in fact allows for the use of mixed methods based on their relevance for addressing research questions, which is essential considering how multi-faceted a solution to alt text barriers needs to be, as discussed throughout the previous chapters. Further, a pragmatist research philosophy is deemed vital for a practical and novel contribution towards addressing alt text barriers, as other philosophies lack the methodological flexibility for designing and evaluating a human-centred practical solution.

7.2.1.2 Approach to theory development: Abduction

Contextualised in pragmatism and driven by gaps in relevant literature and existing solutions to address alt text barriers, this research will use an abductive approach to theory development. Abduction broadly refers to the exploration of a 'surprising' phenomenon to identify patterns and generate themes for informing existing or developing novel theory, which is further tested with the collection and analysis of empirical data (Suddaby, 2006). Abduction has also been defined as an approach to reasoning that seeks a middle ground between deduction (developing new theory by testing existing theory) and induction (developing new theory by collecting and analysing empirical data) (Thompson, 2022). Nonetheless, deductive and inductive approaches to theory development were also considered, but the former was not chosen for not mapping well to the need to gather empirical data from alt text creators and consumers. This choice was further motivated by the gaps in alt text literature about suitability and related solutions, which makes a deductive approach less viable, as it only relies on testing existing theory. An inductive approach that resorts to empirical data without considering existing theory, however, was not selected to avoid abstract answers to the problem posed at the beginning of this research. The abductive approach that will be followed, therefore, will aim for a mixed engagement with existing theory and data collected from empirical studies to explore novel practical solutions to alt text barriers.

7.2.1.3 Methodological choice: Mixed method complex

The chosen philosophy in pragmatism and the adoption of abduction to develop theory allow for the use of both quantitative and qualitative methods. A mixed-methods approach is not only

adequate but essential, as the design and development of a GWAP for authoring and evaluating alt text needs to be informed by qualitative data, i.e., rich, empirical insights from both alt text authors and consumers (RQ2). In later stages, quantitative analysis with the use of statistics will need to be used to evaluate the effectiveness of the output of the GWAP (RQ3), while additional analysis will be required to compare the output of the AI model trained on the GWAP-generated data and the baseline model (RQ4). This highlights the multi-stage structure of this research, where both qualitative and quantitative methods are used as needed to address research questions; thus, the research adopts a complex mixed-methods approach (see Fig. 6).

7.2.1.4 Strategy(ies): Action research

The primary strategy adopted in this research is action research, as the emphasis is on the development of a practical solution in the form of a GWAP for authoring and evaluating alt text. In effect, action research is very apt for the multi-stage structure of this research, as it affords iteration within stages, such as problem understanding, solution implementation, and evaluation. Action research will, nonetheless, be supported by additional strategies at certain stages of the research to achieve specific ends. These include experiment, survey, and archival research strategies at the various stages of the research. The use of archival research involves the analysis of existing documents and is evident in Chapters 1-6, where relevant literature is reviewed (RQ1). Survey and experiment strategies are employed and discussed in Chapter 11, where the evaluation of the performance of AI models is presented, which involves an experiment comparing the performance of trained models with pure image processing (RQ4). Additionally, semi-structured interviews will be used early to gather qualitative insights from alt text authors and consumers to inform the design of the GWAP (see chapter 8) (RQ2). However, these strategies were only adopted as needed at each stage of the research to support an action research strategy, as they do not map well to the iterative nature of this research. The action research strategy at the core of this research is, thus, complemented by further strategies at different stages, reflecting the research's mixed-methods and multi-stage structure.

7.2.1.5 Time horizon: Longitudinal

This research adopts a longitudinal time horizon, as data are collected and analysed at multiple points in time. Importantly, it is clarified that the longitudinal nature of the research relates to the time horizon layer of Saunders, Lewis and Thornhill (2009)'s research onion (see Fig. 6); thus, it does not imply repeated observation of the same participants over time. However, individual studies within this research, such as interviews, gameplay, and surveys, are cross-

sectional, capturing snapshots of user perspectives at specific points in time. Taken together, multiple such snapshots will be collected at different stages, constituting a longitudinal process where each stage informs and builds on one another over the course of this research.

7.2.2 Innermost layer

Following the description of the outer layers of the 'research onion', this section describes the data collection and analysis methods that will be used in this research.

7.2.2.1 Data collection

Whilst this research adopts an abductive approach to theory development, an exploratory study following a qualitative approach through semi-structured interviews with alt text authors (web content creators) and consumers (visually impaired users) will be conducted. Semi-structured interviews are generally preferred in this type of research over other interview formats, such as structured and unstructured interviews, which present notable limitations (Adhabi and Anozie, 2017). Structured interviews lack flexibility for interviewees and are thus ill-advised for a qualitative context, while unstructured interviews compromise the reliability and validity of the data due to the absence of an initial structure in relation to the number, order and nature of the questions (Rashidi *et al.*, 2014). Formatting the interviews as semi-structured is, therefore, important to draw on a core set of questions for reliability and validity of collected data while allowing for adaptations to the natural progression of each conversation.

Moreover, the qualitative data from the interviews will inform the design of a GWAP for alt text annotation and evaluation, which is a novel way to use crowdsourcing to address alt text barriers (RQ3). As discussed in previous chapters, there is a persistent trade-off between suitability of alt text and scalability of the approach in current approaches; therefore, a GWAP is proposed as an implementation that has shown promise for similarly complex tasks and large-scale data collection. Once the GWAP has been developed, crowdsourced data collection will begin by engaging participants with no prior alt text expertise into GWAP gameplay. Then, quantitative data will be collected by former GWAP players through an online survey.

7.2.2.2 Data Analysis

The empirical study proposed in the first stage of this research will involve the analysis of the interview transcripts by the researcher following Braun and Clarke (2019, 2021)'s reflexive thematic analysis six-phase approach, which highlights individual researcher subjectivity as the key resource for knowledge generation, as opposed to other thematic analysis approaches relying on the involvement of multiple coders for bias mitigation (Byrne, 2022). It must be

noted, therefore, that consensus of meaning is not sought in this type of analysis, as opposed to the coding reliability thematic analysis approach (Clarke and Braun, 2013). The latter approach is, in fact, often misinterpreted as a reliability measure, rather than a separate thematic analysis approach (Byrne, 2022). Accordingly, the strategy followed over the course of the six-phases of the reflexive thematic analysis is detailed in Table 6 (see section 8.4). The resulting themes will inform the design of the GWAP, which participants will play during the second user study, generating a human-curated output of alt text descriptions and rating scores.

A mixed-method analysis of this output will then follow, based on crowdsourced annotations, user ratings, and consensus measures, where the output's effectiveness, perceived quality, and consistency will be investigated. This analysis will be preceded by a data cleaning stage and will involve a plurality of statistical tests ranging from descriptive statistics to correlation analyses. Once the cleaned data have been used to train the AI model developed in this thesis, the output of the model will be analysed in comparison with the output of an off-the-shelf version of the same model to assess the effectiveness of training in approximately human-level quality alt text, whilst automating the process. This will be achieved by comparing participants' input on quality perceptions of alt text descriptions generated by the two models, which will be examined both for the overall data and via pair testing of the two models' outputs (RQ4). The tests that will be used for this comparison will be determined once the normality of the distribution of the data gathered through the online survey is evaluated, as parametric statistical tests, for example, are inadequate for non-normally distributed data.

For clarity, both qualitative (Chapter 8) and quantitative (Chapters 10 and 11) strands were analytically integrated in this work as follows:

1. Qualitive insights deriving from interviews with alt text authors and consumers were analysed using Braun and Clarke (2019)'s reflexive thematic analysis approach (Table 6), resulting in six themes and design recommendations for alt text solutions based on these themes (Table 7).
2. These were used to guide the design of the GWAP (Chapter 9), where players generated quantitative data in the form of alt text and rating scores, which were analysed via data cleaning, descriptive statistics, correlation, and semantic similarity (Chapter 10).
3. Finally, the cleaned GWAP data were used to train two proof-of-concept ML models that automatically generated alt text. The evaluation of the performance of these models was achieved through quantitative data analysis using non-parametric statistical tests, due to the data not being normally distributed, as well as through an online survey with former players of the GWAP for comparability.

7.3 Construction of interview and survey questions

The interview questions supporting the semi-structured interviews with visually impaired users (VIUs) and web content creators (WCCs) that were conducted in this research (Chapter 8) were developed based on the research objectives (Chapter 1), and literature on web accessibility and alt text (Chapters 2-4). The initial questions were deliberately broad in focus and open-ended and transitioned to increasingly more specific questions narrowing down from accessibility to alt text. This strategy was adopted to maintain the natural flow of the conversation to allow for a more concrete capturing of individual perspectives and description of personal experiences (Adhabi & Anozie, 2017; Creswell & Poth, 2018). Thus, the semi-structured interview format followed was supported by a script of open-ended questions (see Appendices A and B).

Accordingly, the survey questions supporting the evaluation of the performance of one of the AI models that will be proposed in this research were developed to capture participants' perceptions on the quality of automatically generated alt text descriptions. Each question will ask the participants to rate the suitability of an alt text description for an image in a specific context using a 5-point Likert scale, which is preferred for subjective ratings and enables quantitative analysis (Kusmaryono, Wijayanti, and Maharani, 2022). The use of Likert scale for rating the quality of alt text in particular has been recommended in prior work (Changpinyo *et al.*, 2021; Desai *et al.*, 2021). The same question format will be repeated for several image-context-alt text tuples, while clear instructions on alt text suitability will be provided during the entire survey to minimise straightlining responses (Wang and Hau, 2025). A sample of the question format of the online survey is included in Appendix F.

7.4 Chapter summary

This chapter discussed the methods that will be adopted in this work, whilst it used the 'research onion' as a blueprint for mapping the choice of methods to relevant theory. From outermost to innermost layer (see Fig. 6), the solution to alt text barriers proposed in this thesis will be grounded in pragmatist philosophy, and it will follow an abductive approach to theory development. Therefore, a practical solution to address such barriers will be developed in the form of a GWAP for alt text annotation and evaluation, which literature suggests is most promising yet underexplored in the case of alt text. This research is, thus, underpinned by a complex, longitudinal mixed-method approach, with action research as the primary strategy used, as it maps well to the development of a practical solution and to data being collected and

analysed at multiple stages of the research. Fig. 7 below maps the user studies and evaluation discussed in the next chapters to such methodological decisions, and to the relevant overarching research questions (RQs) (Chapter 1).

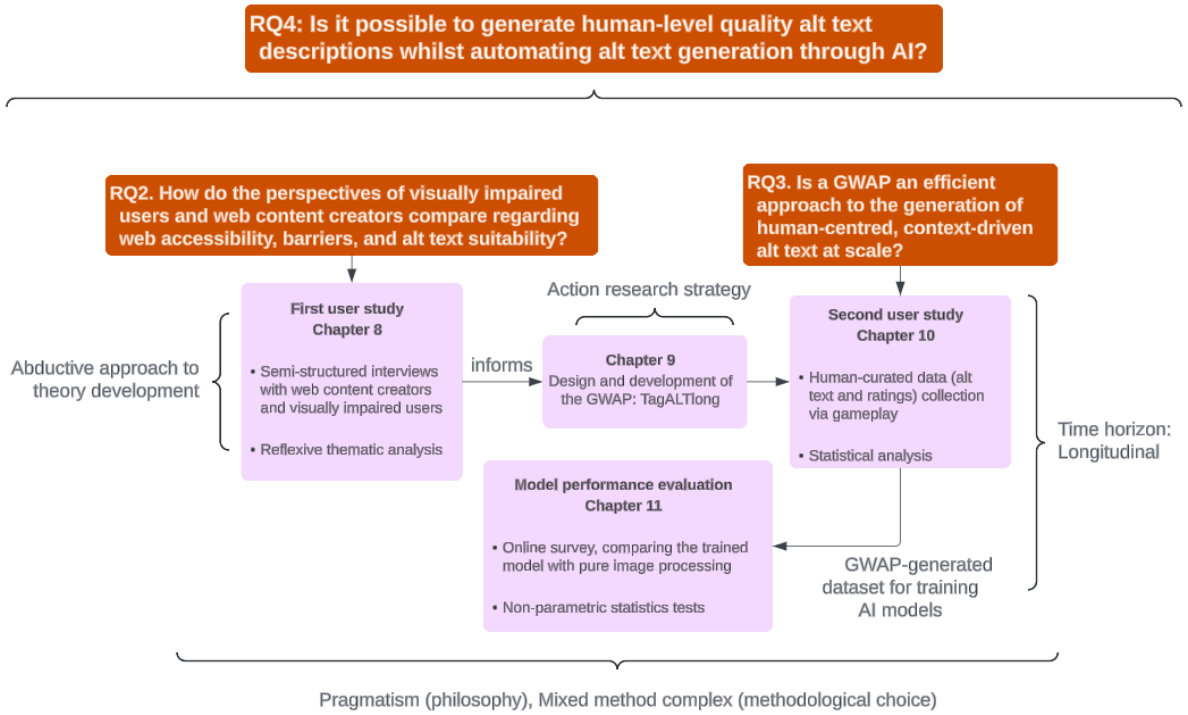


Fig. 7. Diagram mapping user studies to method choices and RQs

First, qualitative data will be collected through semi-structured interviews with web content creators and visually impaired users, and they will be analysed via a reflexive thematic analysis approach to reinforce literature findings with experiential user understandings (RQ2). Taken together, these findings will guide the design of the GWAP solution that will be developed in this thesis, whilst a second user study will be conducted to gather data (alt text descriptions and rating scores) by participants playing the GWAP (RQ3). Once cleaned, the data will be used to train AI models, whose effectiveness will be validated through a twofold evaluation (RQ4). First, a user study via an online survey will be conducted to approximate average human-level quality whilst automating the process. Then, the preservation of context in automatically generated alt text descriptions will be investigated via the binary presence of elements of the proposed context definition (Section 4.3.1). Accordingly, Chapter 8 presents the qualitative study, involving the semi-structured interviews with visually impaired users and web content creators, which is also the first empirical study conducted as part of this thesis.

Chapter 8. First user study: Interviews with visually impaired users and web content creators

8.1 Introduction

The previous chapters highlighted numerous efforts in the literature aimed at improving alt text availability and suitability from an authorship perspective. However, it was also identified that alt text suitability remains largely unaddressed mainly due to the fact that web content creators and web content consumers often have different views on what constitutes suitable alt text in different contexts, whilst available guidelines are inconsistent, ambiguous and do not reflect the plurality of impairments. Accordingly, there is a pressing need to 'clear the air' which will help take a forward leap towards identifying common ground on what constitutes suitable alt text and best ways to achieve this. In response, this chapter presents the first effort to the best of the researcher's knowledge that brings the perceptions of web content authors and visually impaired web content consumers together in an attempt to help bridge this functional perception gap and identify actionable ways forward to improve alt text suitability. It also explores how far scholarly efforts are corroborated by the experiential understandings from both web content creators and visually impaired users; in other words, this work contributes to similar efforts in the literature by providing an empirical account of what web content creators and visually impaired users are also 'saying' as opposed to their experiences of 'doing' in the context of alt text suitability. The findings of this study will further be used to inform the proposed solution (see next chapter). This study addressed three sub-research questions (S-RQs):

- S-RQ1. What are the perceptions of web content creators on the accessibility of the web through screen readers against visually impaired users' web navigation experiences?
- S-RQ2. What are the perceptions of web content creators on WCAG against those of visually impaired users?
- S-RQ3. What makes alt text suitable according to both visually impaired users and web content creators?

In accordance with the methodological framework described in the previous chapter, the first user study conducted in this thesis is presented. The study follows an exploratory, qualitative approach through semi-structured online interviews, followed by reflexive thematic analysis. Accordingly, this chapter presents the participants, the study protocol and the data collection and analysis approaches, resulting in the generation of a set of six themes, which are used to propose much-needed recommendations for authoring suitable alt text.

8.2 Participants and recruitment

In total, 22 participants (11 web content creators (WCC) and 11 visually impaired users (VIU)) were recruited and interviewed from January to March 2024. Six of the participants identified as male and five as female in the former group, whilst eight identified as male and three as female in the latter. The mean age across both groups was 44 years (range 22-70; SD 14). Tables 4 and 5 provide an overview of the two participant groups. The specific inclusion criteria were broad by design to recruit a diverse sample of web content creators and visually impaired users, including (1) being at least 18 years old at the time of the interview, (2) have minimum two years of experience with creating and/or evaluating web content, and (3) some experience with creating accessible web content was desirable, particularly writing and/or evaluating alt text descriptions; for the former. Similarly, the inclusion criteria for the latter included (1) being at least 18 years old at the time of the interview, (2) being a frequent user of the Web, and (3) use or having used screen readers to navigate the Web. The exclusion criteria for both groups were (1) do not speak or understand English and (2) not able to provide consent independently. As such, all participants were fluent in English. All participants were recruited from relevant institutions and organizations, including the Royal National Institute of Blind People (UK),[5] WebAIM,[6] AbilityNet,[7] the National Federation of the Blind (Greece),[8] Silktide,[9] KreativeInc Agency Ltd,[10] and Scope.[11] It was requested that they share the call for participation with their members through internal mailing lists. Snowball sampling was then used until saturation was achieved. Participants who were interested in participating contacted the researcher, and if they qualified based on the inclusion and exclusion criteria, they were then handed a participant information sheet and a consent form to sign before proceeding to scheduling an interview. Potential participants were also informed that interviews would be recorded. The ethics protocol was approved by the institutional Research Ethics Committee (Ref: 41665-LR-Jun/2023- 45191-3).

Table 4. Visually impaired users and self-reported experiences

ID	Age range	Gender	Visual Impairment	Yrs. of web Use	Yrs. of Screen Reader Use
VIU1	18 – 24	Male	Blindness	15	15

5 https://www.rnib.org.uk
6 https://webaim.org
7 https://abilitynet.org.uk
8 https://www.eoty.gr
9 https://silktide.com
10 https://kreativeincagency.co.uk
11 https://www.scope.org.uk

ID	Age range	Gender	Visual Impairment	Yrs. of web Use	Yrs. of Screen Reader Use
VIU2	45 – 54	Male	Retinitis pigmentosa	30	27
VIU3	55 – 64	Female	Retinopathy of prematurity	30	37
VIU4	18 – 24	Female	Severe sight impairment (Registered blindness)	6	6
VIU5	35 – 44	Female	Blindness	24	26
VIU6	25 – 34	Male	Severe sight impairment (Registered blindness)	10	10
VIU7	35 – 44	Male	Blindness since birth	25	25
VIU8	45 – 54	Male	Blindness NLP (No light perception)	26	28
VIU9	65+	Male	Uveitis (Registered severe visual impairment – Blindness)	30	18
VIU10	45 - 54	Male	Blindness	30	35
VIU11	45 - 54	Male	Blindness	30	30

Table 5. Web content creators and self-reported experiences

ID	Age range	Gender	Job Title	Web Experience (in yrs)	Accessibility Experience (in yrs)
WCC1	65+	Male	Web Accessibility Consultant	28	20
WCC2	55 – 64	Female	Accessibility Coordinator	3	3
WCC3	45 – 54	Male	Digital Delivery Manager and Digital Accessibility Lead	25	5
WCC4	25 – 34	Male	Accessibility Engineer and Consultant	7	7
WCC5	45 – 54	Female	Senior Accessibility Engineer	12	12
WCC6	55 – 64	Female	Web Accessibility Consultant	14	4
WCC7	18 – 24	Male	Digital Media Developer	4	0
WCC8	25 – 34	Female	Digital Accessibility Specialist	3	3
WCC9	35 - 44	Female	User Experience (UX) Designer	10	5
WCC10	45 - 54	Male	Business Owner	18	18
WCC11	35 - 44	Male	Principal Engineer	22	16

As can be inferred from the above tables, inclusion criteria were deliberately broad to capture diverse viewpoints within each user group. Indicatively, visual impairments for the VIU participants (Table 4) ranged from blindness to retinitis pigmentosa, while screen reader use ranged from 6 to 37 years. This assured representation across the spectrum of visual impairment in the sample of 11 VIU participants, while the WCC participants (Table 5) came from diverse professional roles (e.g., digital media developer and digital accessibility specialist), with their experience in accessibility ranging from 0 to 20 years, thereby capturing both novice and expert views on accessibility and alt text. A further choice during recruitment was the need to contact participants through institutions, such as WebAIM, AbilityNet and Silktide, to ensure that the sample of participants had active engagement with accessibility issues and barriers. The aim was therefore for perspective richness on topics, such as web navigation via screen readers, alt text and suitability of alt text, rather than generalizability. Snowball sampling then continued and data saturation was observed after eight interviews in each of the groups. Finally, the sample size is also consistent with peer studies (e.g., Muehlbradt and Kane, 2022; Mack et al., 2021; Aizpurua et al., 2016; Lee and Ashok, 2022).

8.3 Interview protocol

The interviews took place online using *Zoom* or *Microsoft Teams* video conferencing software and oral or written informed consent was obtained from each participant beforehand. A semi-structured interview format was followed using an open-ended questions script by design (see Appendices A and B) to encourage participants to share their personal experiences and insights. For web content creators, the interviews included questions about their experience with accessible web content creation, related accessibility barriers, and their experience authoring or evaluating alt text, as well as their expectations for its suitability. Visually impaired users were asked questions about their web browsing experience and associated barriers, screen reader usage experience, as well as their experience with alt text and their expectations with regard to its suitability. Each interview lasted between 45 and 100 minutes, and they were video recorded while the researcher was also taking notes by hand.

Specifically, the interview process was composed of four parts for both participant groups. For visually impaired users, demographic and general questions were first asked to elaborate on their web navigation experience. In the second part, they were asked more specifically about screen readers and web accessibility barriers. Finally, in the third part of the interview, they were asked about their experience and expectations with alt text. The interview was then concluded with a website browsing task, where they were asked to browse the inaccessible and

accessible versions of the 'News Page' from W3C's Demo[12] and, as such, share their opinion in relation to alt text. For web content creators, they were first asked demographic and general questions about their years of experience in web content creation and related accessibility efforts. In the second part, they were asked about web accessibility barriers, particularly those related to the use of screen readers, and what do they do to deal with such barriers, as well as their familiarisation with WCAG. Then, in the third part of the interview, they were asked more specifically about alt text, including their experience in authoring or evaluating alt text, as well as their perceptions as regards its suitability. Finally, the interviews with web content creators were also concluded with the same website browsing task, where they were asked about the sufficiency of the 'accessible' version, not least in relation to alt text.

It must be noted at this stage that the website browsing task was part of the interview and not a separate task, thus the generated qualitative insights were formulated into a single data set, which were then captured in the themes resulting from the data analysis process explained in the next section. Also, whilst the intention was to follow the above question flow, this was occasionally altered to accommodate each discussion and how it progressed.

8.4 Data analysis

In total, close to 25 hours of interviews were recorded and transcribed. Interview transcripts were analysed by the researcher following Braun and Clarke (2019, 2021)'s reflexive thematic analysis six-phase approach, the adequacy of which was discussed in the previous chapter. A member of the supervisory team was asked to independently sense-check the themes and narrative at the end of the six phases of the analysis outlined in Table 6. It must be noted that the involvement of more than one researcher was made in accordance with principles of the reflexive approach aiming for increased nuance of meaning, rather than achieving consensus of meaning, as it is common in the coding reliability thematic analysis approach, which is often misinterpreted as a reliability measure, rather than a separate thematic analysis approach (see section 7.2.2).

Trustworthiness and reliability criteria that were consistent with the reflexive approach were achieved through ensuring a rich description of the analysis process and by including plentiful descriptions of participant quotes (Nowell *et al.*, 2017). This resulted in a list of close to 400

[12] https://www.w3.org/WAI/demos/bad/Overview.html

codes, which were later revised by the researcher through the phases discussed in Table 6 below to arrive at the broader themes presented in the findings.

Table 6. Reflexive thematic analysis phases and descriptions

Phases of Reflexive Thematic Analysis	Phase Adaptation Description and Trustworthiness
Phase one: Familiarization with the data	First, the researcher revisited the physical notes he had taken while recording the interviews and then transcribed the data in *Microsoft Excel* spreadsheets after listening to the recordings for a general understanding and engagement with the data corpus as a whole.
Phase two: Generating initial codes	The researcher generated initial codes for the entire data corpus to avoid missing links between data items (Braun and Clarke, 2006). Both latent and semantic coding were used, with no attempt to prioritise one over the other on any given occasion to ensure interpretation of both participant-communicated and researcher-interpreted meaning (Patton, 1990).
Phase three: Generating themes	Next, the researcher compiled the full list of codes in search of shared meaning between the codes to generate themes and their respective subthemes (Braun and Clarke, 2016).
Phase four: Reviewing potential themes	The researcher aimed to finalise the list of themes using Patton (1990)'s dual criteria, i.e., internal homogeneity within the themes and external homogeneity among the themes. Six themes were conceived in the analysis of the entire data set.
Phase five: Defining and naming theme	The researcher then revisited and refined the names of the themes to divert from names that wholly described each theme to captivating names highlighting one important aspect of the theme in question, and that can later be understood in detail via an analytic narrative (Braun and Clarke, 2021).
Phase six: Producing the report	Finally, the researcher conceived an *analytic* narrative, consistently with Braun and Clarke (2019)'s instructions, that includes data extracts scrutinized in relation to theory and the S-RQs as and when they are reported. The narrative was reviewed by one more member of the supervisory team for coherence and trustworthiness.

For completeness, and deferring to Byrne (2022)'s guide on the reportability of the above six phases, a worked example relating to the three themes generated for WCC participants follows prior to the findings to ensure transparency of the theme generation process. Deferring first to 'Phase one: Familiarization with the data' (Table 6) manual transcription of the data after actively listening to the recordings of the interviews took place by the researcher to allow for a general understanding and engagement with the data corpus as a whole. Before moving on to phase two, it is important to clarify the following:

> *"Data corpus: All data collected for a particular research project.*
> *Data set: All the data from the corpus that are being used for a particular analysis.*
> *Data item: Each individual piece of data collected.*

> *Data extract: An individual coded chunk of data, which has been identified within, and extracted from, a data item."*
> (Braun and Clarke, 2006, p. 79)

In phase two, the generation of initial codes, namely "entities that capture (at least) one observation, display (usually just) one facet; themes, in contrast, … capture multiple observations or facets" (Braun and Clarke, 2021, p. 340), took place, and a short excerpt of the coding process for a data extract from a Web content creator, with each code included in a parenthesis right after the portion of text that informed its generation, follows:

> "There was always a drive to make things accessible (drive for accessibility), even though nobody knew what accessibility really is (accessibility ignorance) and I don't think there were any guidelines at this point (lack of guidelines). It isn't so much sticking to guidance (beyond guidelines) but it's more user testing (user testing necessity) and you know that we prototype stuff. It's important to initially understand why accessibility is needed, why are we doing what we are doing (understand accessibility benefits), and get people talking about it to normalize the attitude towards it (normalize attitudes towards accessibility), because really making stuff accessible especially on the Web is… just doing your job properly! (content creators' responsibility to create only accessible stuff). My experience is that stuff gets de-prioritized (accessibility de-prioritization) and that attitude on accessibility should be turned around (need to shift current accessibility attitudes). That's one of the biggest hurdles; changing people's attitudes and also getting senior leadership on board (need for senior leadership to support accessibility)." [WCC3]

It must first be noted that the entire data corpus was coded before moving on to the generation of initial themes, to avoid missing links between data items (Braun and Clarke, 2006), and the above coded extract is only included as an example. Accordingly, and as derives from the above example, both latent and semantic coding, namely coding that involves and does not involve interpretation, respectively, were used, with no attempt to prioritise one over the other on any given occasion (see Table 6). Following on from phase two, the full list of codes is compiled in search of shared meaning between the codes for the generation of themes and their respective subthemes; indicatively:

> "*A subtheme exists 'underneath' the umbrella of a theme. It shares the same central organising concept as the them, but focuses on one notable specific element … Through naming and analysing a specific subtheme, that aspect of the theme becomes more salient."* (Braun and Clarke, 2016, para. 1)

Accordingly and for the sake of example, a thematic map of the initial list of themes and their respective subthemes for WCCs can be seen in Fig. 8 below.

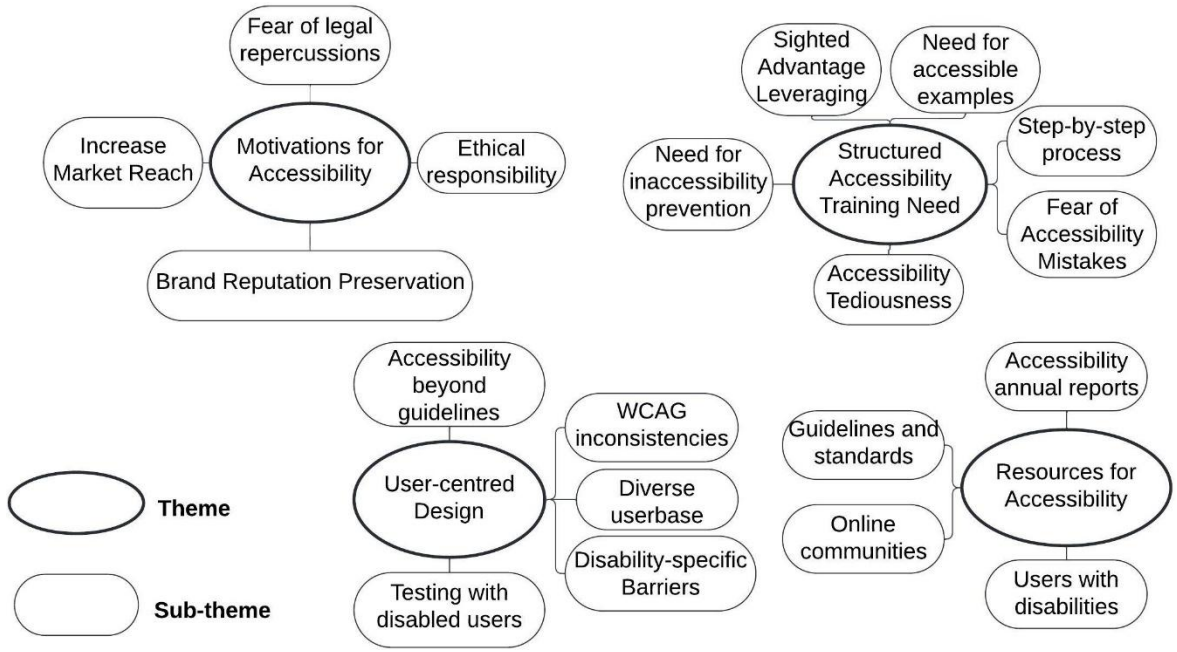


Fig. 8. Initial thematic map for WCCs indicating four candidate themes and their subthemes

Deferring to the above figure, concurrent narratives deriving from the compiled list of codes for the entire data corpus for WCCs were grouped under subthemes, such as ‘fear of legal repercussions’, ‘ethical responsibility’, ‘brand reputation preservation’, and ‘increase market reach’, and the initial name ‘motivations for accessibility’ was given to the theme encompassing them, owing to their close relation with what motivates WCCs to engage with web accessibility. Similarly, ‘structured accessibility training need’ was the initially suggested theme name for capturing the ‘need for inaccessibility prevention’, ‘step-by-step process’, ‘fear of accessibility mistakes’, ‘accessibility tediousness’, ‘sighted advantage leveraging’, and the ‘need for accessible examples’, which are connected to reasons for not engaging with training in web accessibility and suggestions for improving existing training means. ‘Resources for accessibility’ includes subthemes on where WCCs turn to for increasing their understanding of web accessibility, that is, ‘guidelines and standards’, ‘online communities’, ‘users with disabilities’, and ‘accessibility annual reports’. Finally, ‘user-centred design’ encompasses these narratives that highlighted unaddressed needs in designing for accessibility, that is, ‘accessibility beyond guidelines, ‘WCAG inconsistencies’, ‘diverse userbase’, ‘testing with disabled users’, and ‘disability-specific barriers’, which are in line with user-centred design practices.

In phases four and five (Table 6), the list of themes was revisited and refined, as were the names of the themes to divert from names that wholly describe the theme in question to captivating names highlighting one important facet of each theme, and that can later be

understood in detail via an analytic narrative (Braun and Clarke, 2021). Accordingly, the finalized list of themes and subthemes for WCCs can be seen in Fig. 9 below.

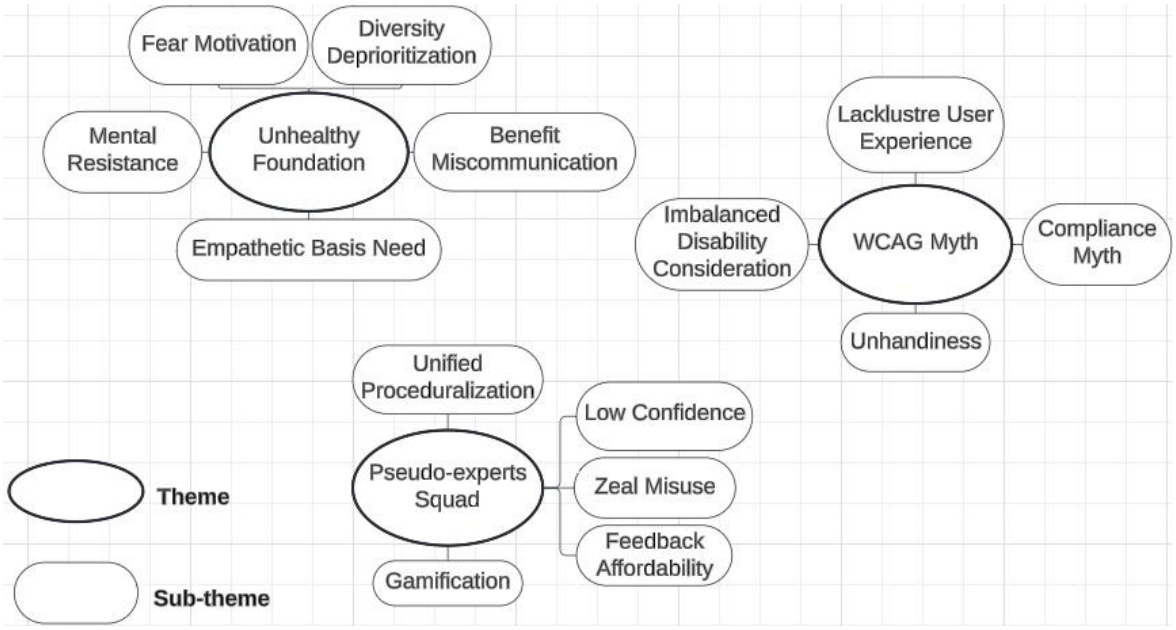


Fig. 9. Finalized thematic map for WCCs demonstrating three themes and their subthemes

Deferring to the above figure, the motivations for accessibility were reviewed in more detail and revealed that they were pertaining to an ‘unhealthy foundation’ as an overarching theme, as shown by the subthemes ‘mental resistance’, ‘fear motivation’, diversity deprioritization’, ‘benefit miscommunication’, and ‘empathetic basis need’. Similarly, the focus on WCAG became evident, with particular focus on how they are misinterpreted as standards to conform with while they are a set of guidelines; hence, the theme ‘WCAG myth’, informed by the subthemes ‘compliance myth’, ‘unhandiness’, ‘lacklustre user experience’, and ‘imbalanced disability consideration’. Finally, ‘pseudo-experts squad’ captures the user-centred design needs mentioned in the initial list of themes and extends it to also include how web accessibility training is an ongoing process and the afore-mentioned high variability of expertise in web accessibility. Therefore, the theme draws from both ‘user-centred design’ and ‘structured accessibility training need’ themes from the first list, and includes the following subthemes: ‘unified proceduralization’, ‘gamification’, ‘low confidence’, ‘feedback affordability’, and ‘zeal misuse’. Importantly, there is no correct number of themes in reflexive thematic analysis, with too many themes posing coherence-related risks and too few themes posing depth- and breadth-related risks (Byrne, 2022). Six themes were conceived in our analysis of the entire data set, aligning with Lichtman (2013)’s rule of thumb of not exceeding five to seven themes per data set. Accordingly, Patton (1990)’s dual criteria, namely internal homogeneity within the themes and external homogeneity among the themes, were used to

arrive at both final sets of themes. The use of these criteria becomes more clear in the following section, where, consistently with Braun and Clarke (2019)'s instructions, an *analytic* narrative is presented including data extracts scrutinized in relation to theory and the RQs as and when they are reported, which is considered the most optimal way for disseminating results from reflexive thematic analyses, as opposed to the typical split between a 'Results' and a 'Discussion' section (Clarke and Braun, 2013; Terry *et al.*, 2017).

8.5 Findings

In this section, the analysis of the themes using key data extracts is reported, highlighting each theme's unique nuances and anchoring them to the scholarly field and the S-RQs. Finally, to the best of the researcher's knowledge, the first much-needed set of recommendations on alt text suitability and trainability is presented that takes into account the needs of both visually impaired users and web content creators.

8.5.1 Web content creator perceptions

The findings from the semi-structured interviews with the 11 web content creators are first presented focusing on their perceptions on web accessibility for screen readers, WCAG conformance, and alt text suitability, as well as reasons for the deprioritisation of web accessibility. Accordingly, a set of trainability recommendations for accessibility-related training has been established and presented.

8.5.1.1 Unhealthy foundation (S-RQ1, S-RQ3)

A mental resistance on the part of WCCs to engage with web accessibility was a key factor among WCCs indicating that there seems to be an unhealthy foundation from the outset. WCC10 notably emphasised that "Once you get people's mind changed everything else falls into place!," which is in line with past evidence on the low allocation of resources towards accessibility and the reluctance of WCCs to engage with web accessibility-related training that points to such an unhealthy foundation (Williams et al., 2022). WCCs appear to be aware of the numerous benefits that web accessibility offers highlighting that "There's definitely business benefits to it. You're missing out on hundreds of millions of potential customers for example … They also have billions of dollars' worth of money to spend that you're also then missing out on … .Also, there's the legal case for it, which I think is why we've seen a lot of people and companies starting to care about, because they're getting sued … And I think related to that, which I kind of said earlier, when you create accessible web experiences from the

beginning, you're also making it even—you may have unexpected benefits or making it accessible and usable for other people in ways that you haven't foreseen." [WCC9].

In particular, a decreased market reach due to low accessibility efforts is in line with previous evidence on missing out on growing, aging, and impaired user markets (Waller *et al.*, 2015; Moreno *et al.*, 2019), whilst in terms of inclusivity it has been shown that it can both increase brand reputation and improve web navigation for all users, owing to inclusive web products being 35% more usable by everyone (Clark, 2001). Nevertheless, these benefits are often miscommunicated and WCCs are instead being warned about potential legal repercussions if they do not focus their efforts on accessibility, as WCC2 indicated that "The main benefit is that you're less likely to be sued. Well, if I'm being realistic that is why. I think that's why they created these new policies that we had to follow, but of course they were creating policies when we didn't have people on staff with the expertise to meet these policies." This extract is very telling in relation to where WCCs' mental resistance stems from indicating that the main benefit that is communicated to them relates to fear motivation. Indeed, past similar work corroborates that avoidance of legal repercussions was by far the most cited motivational factor for WCCs to engage with accessibility (Open Inclusion, 2022). Accordingly, the above further supports that not facing legal repercussions is the main consideration of businesses, especially considering brand reputation which was highlighted as another key motivational factor (Kaur and Kumar, 2015a). Nevertheless, the interviews surmised that intrinsic motivational factors should instead be emphasized with WCC11 indicating that "In general, it's a legal requirement to make accessible websites.—Just in general, it's like, you know, if I can spend that hour or whatever making this work for everyone, why wouldn't I?"

Unsurprisingly, the legal case for web accessibility is also critiqued as an unhealthy approach to highlight benefits that can better motivate WCCs. Mott *et al.* (2019) suggest that if WCCs were to adopt a more positive mentality towards web accessibility, then there would be more flexibility towards overall user needs. This was supported by WCC3 who discussed that "It's important to initially understand why accessibility is needed … and get people talking about it. To normalise attitudes towards it, because really making stuff accessible, especially on the Web, is just doing your job properly. My experience is that stuff gets deprioritised and that attitudes on accessibility should be turned around. It's one of the biggest hurdles: changing people's attitudes." This further highlights the connection between the mentality resistance and the miscommunication of benefits of web accessibility, acknowledging both how accessibility is deprioritised and how it is the responsibility of WCCs to deliver accessibility.

Past research is in fact in line with accessibility being an integral responsibility of the WCC role (Power and Petrie, 2007; Crespo, Espada and Burgos, 2016), however, overcoming WCCs' mentality resistance towards inclusivity appears to be the most persistent challenge, which agrees well with Nedelkina (2022)'s notion that WCCs often prefer to rely on stereotypes and their own assumptions about the web navigation experience of users with disabilities. Instead, as WCC8 put forward "I think it's that like push to tell people: "C'mon guys do this accessibly!' … because they might be missing on something that can only be experienced by someone who uses a screen reader frequently. I think one of the biggest challenges at the moment is actually when people talk about: "We're doing accessibility and what they mean is that they have checked it with a screen reader … They don't think of other things like color contrast or that not everyone who is visually impaired accesses things in the same ways." In an alt text context, for example, WCC10 highlights one of the most encountered mishaps on the web stating "'An image of a cat,' because alt text describes an image, so you don't say: 'An image of,' you say: 'A cat.' It [the screen reader] knows it's an image or graphic, so you don't say: 'A photo of,' 'An image of;' you just say what it is." Similar mishaps related to the length and the language alt text is authored in are addressed by WCC9, stating that "If you use too much information it might be necessary and it might be annoying to the people using screen readers as well, so the length of your alt text definitely needs to be considered; not too long but also not too short; that's not useful either … They need concepts broken down into plain language; they shouldn't be reading like all these things being written on a graduate level."

Evidently, however, current accessible web design efforts focus more closely on specific impairments and, as such, deprioritise diversity, which aligns with Aizpurua, Harper and Vigo (2016)'s previously identified functional gap between how inaccessibility is perceived and how it is experienced. This is reminiscent of the large body of accessibility literature explaining that guidelines and scholarly efforts are overfocused on blindness (Friedman and Bryen, 2007; Miranda and Araujo, 2022). The findings of this work on the other hand stress the need to foster empathy towards inaccessible web navigation experiences to transition to understanding and designing for accessibility. Vollenwyder *et al.* (2023) have, in fact, recently shown how WCCs' motivation to engage with web accessibility increases when they are first given a chance to relate to what inaccessible web navigation feels like. It can thus be conjectured that offering a glimpse into inaccessibility is a promising way against the unhealthy foundation that WCCs' mental resistance to engage with web accessibility stems from. Importantly, it can also be noted that sentiments discussed in this section aligned with WCCs' profiles (Table 5); i.e., WCC3's three year of experience in accessibility appears to explain exposure to fear-based motivation

and the unhealthy motivation, while WCC11 advocates intrinsic motivation and caring for the end-user, which aligns with their 11 years of accessibility experience.

8.5.1.2 WCAG myth (S-RQ2, S-RQ3)

Chapter 2 also highlighted that WCCs typically over rely on the WCAG, which are principally meant to guide rather than dictate how to create accessible web products. WCC5 in particular confirmed this: "What is WCAG? They're actually guidelines—Insert joke from the Pirates of the Caribbean: 'The Code is more what you'd call guidelines than actual rules!'," which is in line with past research calling attention to the insufficiency of WCAG to fully capture web accessibility (McCarthy and Swierenga, 2010; Crespo et al., 2016). Other participants appeared to be in agreement with this notion, with WCC3 stating that WCAG is " … a piece of documentation that is widely misunderstood. A lot of government and regulators will point to WCAG as a standard while it's not a standard, it's a guideline, and a standard is something that you have to meet hence people talk about compliance all the time, but actually a guideline is: "Broadly speaking, in this situation you need to have a thing that works and looks like this." It's not the law, so literally, it's not the law. And people often go: "Oh, do you meet the standard?"" This last comment further highlights that WCAG are far too often misinterpreted as standards rather than guidelines, and this appears to be the main reason that WCCs abide by the WCAG conformance logic, which more closely relates to standards.

It is in fact evident that a lot of academic scrutiny has gone into WCAG conformance (Cooper, 2016; Lengua, Rubano and Vitali, 2022), which highlights that WCAG conformance is often inadequate considering that they are guidelines that are often misinterpreted as rules. This view was not shared by all participants with WCC11 stating "I view the standards as tools really. The ultimate goal here is not to conform to a document. It's to create a good user experience and if it came down to following the rules in a doc… I would rather create a good user experience." Interestingly, however, recent work by McCall and Chagnon (2022) showed that usability and user experience are all but ignored by WCAG conformance, with the former especially being considered a prerequisite to more holistically address user experience in a web context (Gartland *et al.*, 2022). This finding was confirmed by participants with WCC6 highlighting that "A website that conforms to WCAG is not necessarily a user-friendly website. Just building a website to WCAG regulations and then assessing it like that is like assessing a meal by the ingredients and not by the taste of it." However, it has to be noted that previous work suggested that WCAG conformance should be a first, albeit ironclad, step towards the creation of accessible web products (Dobransky and Hargittai, 2016; Power, Cairns and Barlet,

2018), which the authors are in agreement with in the efforts to address issues with an unhealthy foundation (see section 8.5.1.1).

Finally, a participant (WCC8) emphasised that WCAG are not particularly helpful as a comprehensive resource to guide accessible web design decisions stating that "There are initiatives to turn the language that the WCAG guidelines are written in into plain English. They are a nightmare!," which further fosters an unhealthy foundation. In fact, the complexity of WCAG is not new (Spyridonis, Daylamani-Zad and Paraskevopoulos, 2017), which led to various efforts in the literature to increase the motivation of WCCs to engage with the WCAG (Chatziemmanouil and Katsanos, 2024; Lorgat, Paredes and Rocha, 2024). More specifically in relation to alt text, WCC4 highlights an important distinction between alt text and plain text; indicatively, "Putting all that huge information as a text alternative is very bad, because if it were text, the screen reader would have the ability to go line by line, and if they do not understand, they can go back to the previous line and they can go back to the next one later, but the text alternative will get announced all at once, so the screen reader will not have the ability to, okay, I want to hear again, this particular part of it. They won't have this ability." This ironclad distinction between alt text and plain text with regards to how they are being treated by screen readers is corroborated by VIUs (see section 8.5.2). To the best of our knowledge, however, it has never been formalised in a scholarly context or in well-acclaimed web accessibility guidelines, although it has been identified and reported in certain accessibility resources (Accessibility for Ontarians with Disabilities Act, 2023). Accordingly, this explains the shared expectations on alt text for graphs, with WCC3 emphasizing "You've got to have a brief description of the data represented underneath in a couple of sentences. Also, you'd link back to the source data. The user would want to find out what it was from some other place, which'd be a broader piece of research, but from that page they can get a high-level understanding of what that thing represents. Let's say: "This graph shows blah blah blah, the summary of which is this, and that's it." Taken together, alt text for graphs is best approached with a brief description that includes the type of the graph and any conclusion that can be drawn from it, as well as information about where a detailed description in plain text can be found.

8.5.1.3 Pseudo-experts squad (S-RQ1, S-RQ3)

The issues identified in the previous sections are exacerbated by the reported low relevance of web accessibility "expertise", which has been shown to vary in multiple occasions (Petrie *et al.*, 2011). The importance of the variability and diversity of perceptions with respect to web accessibility expertise was highlighted by WCC5 stating that "If you ask: 'Is this an accessible

thing?,' and you ask five different accessibility experts, you're gonna get six different opinions." This is very evident in the context of alt text, where some WCCs advise that no image is purely decorative, namely images that add nothing beyond visual aesthetics to a Webpage: "Decorative images enhance the appreciation of a Webpage. Images of all kinds do. So, I think pretty much all images should have alt text. It's back to poetry, yeah?" [WCC10], while others highlight the need for such images to be marked as decorative, so that screen readers skip them during navigation: "Don't be afraid to mark things as decorative, you see far too much alt text on stuff that's decorative and I think people are worried that they are gonna get it wrong." [WCC8]. Drawing on the latter extract, it can be surmised that in the absence of a healthy foundation, adequate support is not in place for WCCs to confidently decide on whether images should or not be marked as decorative in different contexts.

Moreover, a different participant (WCC3) emphasised that building a healthier foundation for engaging WCCs with web accessibility is imperative and that web accessibility guidelines should only complement such a foundation as support tools:" … you wanna get people to understand why they're doing it and who they're doing it for…I don't wanna say you don't need the guidance, but the guidance becomes a support. Making sure that the right support is in place, so that they're allowed to make mistakes. If you support people when they didn't do something right, and they should have done it, then the next time they do it, they'll do it right. And also, if you make people not afraid to ask questions." Importantly, this highlights a low confidence of WCCs to make accessibility-related decisions out of fear of making mistakes, especially when knowing the impact of such decisions to visually impaired users.

The need for confidence in one's own ability to create web products that are accessible is indeed emphasised in WCC9's comment: "I always kind of doubt myself because I don't know how I compare to other people's skillset, but I feel confident. I have done quite a bit of reading and I've applied things to the work I do, but there's always more to learn for sure." In addition, this participant stresses the need for good and bad examples of accessibility practice, as well as a way to assess one's understanding: "In addition to the guidelines if there was more examples and I feel there's never enough examples. I want multiple examples so I can understand and, you know, in different contexts what is a good example of alt text and what is sufficient and maybe also examples of what is bad alt text, so the more examples you can give the more it makes sense and then on top of that if there was some type of tool or a quiz that you could take that you're maybe given a photo and you have to generate the alt text so that it can somehow be graded."

However, the need for proceduralising specific web accessibility tasks, such as alt text authorship, is not encompassed by existing guidelines. Interestingly, this was picked up by one of the participants (WCC11) who suggested that what they " … would like to see is a tool for developers where they experience the web as a text adventure like a forest and here's a well, using the accessibility tree, you know, like can you navigate it using that kind of navigation? And I think that would sort of build empathy and also, yeah, surface accessibility challenges." In a similar vein, WCC6 touched upon the need for the learning process to become more informed and constructive, suggesting providing " … feedback to make some more changes and make the Website even better. It's that openness to learn." Both past research (Abuaddous, Jali and Basir, 2016; Lengua, Rubano and Vitali, 2022) and findings in this work suggest that the mental resistance to engage WCCs with accessibility is the greatest challenge; thus, it is imperative that the right support is in place to leverage their zeal when WCCs are engaged, so that such zeal is not misused.

Finally, the above comment further highlights the need to afford opportunities for visually impaired users to reach out to WCCs about anything that they have found to be inaccessible on the Web. This has in fact been recently suggested by Loseby (2023) and is in line with a recent user survey revealing that 67% of users seldom or never reach out to WCCs about encountered barriers, but it remains unclear if the websites allowed for them to reach out in the first place (WebAIM, 2024a). Reaching out to WCCs is therefore essential, as it has been advocated that the only true experts in accessibility are those who experience inaccessibility (Vollenwyder *et al.*, 2020; Muehlbradt and Kane, 2022). Importantly, it can also be noted that sentiments in this section are again tied to the profiles of WCCs (Table 5); i.e., WCC5 speaks of variability in the perspectives of WCCs on what is accessible, while WCC7 has not encountered such variability, which reflects their respective experience in accessibility (12 and 0 years, respectively).

8.5.1.4 Trainability recommendations per WCCs

This section discussed the perceptions of WCCs in relation to web accessibility, not least in relation to screen readers and alt text. WCCs emphasise the need to build a healthier foundation for engaging with web accessibility-related training, as current motivational factors and official guidelines, such as WCAG, are insufficient and are being misinterpreted, respectively. Taken together, they create a mental resistance which our experienced WCCs deem as accessibility efforts' worst enemy. As such, the findings so far point to the need for accessibility-related training that:

- Is structured: Coaxing WCCs into understanding how to deliver accessibility, rather

than only overwhelming them with complex and gargantuan documentation, such as WCAG (Sections 8.5.1.1 and 8.5.1.2).

- Is example-driven: Allows for the use of good/bad examples, e.g., suitable/unsuitable alt text in different contexts, to coax WCCs into accessibility (Section 8.5.1.3).
- Is appreciative of reasons that demotivate WCCs to engage with training: There is no 'one-size-fits-all' to accessibility to alleviate WCCs' atelophobia as regards time and cost-of-error (Sections 8.4.1.1 and 8.4.1.3).
- Is inclusive of reaching out opportunities: Allows for VIUs to reach out to WCCs when they encounter barriers, as even when accessibility expertise is high, it is important to respect that VIUs are the only ones who can tell whether a website is accessible, useable and/or user-friendly to navigate via a screen reader (Sections 8.4.1.2 and 8.4.1.3).

8.5.2 Visually impaired user perceptions

Following on from the findings from the interviews with WCCs, the results from the semi-structured interviews with the 11 visually impaired users are presented next focusing on their experiential understanding of web navigation via screen readers, their preferred role in relation to the authorship of alt text, and their perceptions of what makes alt text suitable. As in the previous section, a second set of trainability recommendations for accessibility-related training has been established.

8.5.2.1 Coin flipping (S-RQ1, S-RQ3)

Unsurprisingly, VIUs appear to have low expectations on web accessibility, not least in relation to alt text availability and suitability, as they are typically used to no alt text being available. VIU4's comment on how web navigation using screen readers resembles a "coin flip," i.e., the result is either an accessible or inaccessible website, is alarming: "It's about how lucky you get. If you're lucky you get a description and you can get an idea. Sometimes you might not get a description at all, or the description might not be very clear. It's all about luck." This aligns well with recent findings on the minuscule (2.2%) decrease in unsuitable alt text over the last five years compared to a general increase in alt text provision (Muehlbradt and Kane, 2022; WebAIM, 2024b). This is supported by a different participant (VIU6) who stated that "I find navigation a bit difficult. I am worried about whether something is accessible more than whether it's usable or enjoyable. There isn't a lot out there, because I don't really expect the websites to have alternative descriptions for example, because a lot of them don't. It's more like if they got a description then that's great, but there's probably not gonna be a description.

Feels like you are stuck. It's like you can only go so far down until you get stuck." This latest comment further highlights a worry of ending up on the wrong side of the coin flip, which has them disregarding usability and user experience; interestingly, this is reminiscent of this work's findings about accessibility through WCAG conformance (see section 8.5.1.2). Zong *et al.* (2022) have in fact implied that web accessibility-related decisions are made only by WCCs, and Miranda and Araujo (2022) have recently shown that such decisions are typically limited to WCAG conformance.

The above participants' comments highlighted key aspects of the VIUs' web navigation experience that seem to be very much aligned with this work's findings on how WCCs typically approach web accessibility. Their concerns and low expectations extend to other types of media too, with VIU1 stating "I've only experienced alt text for images and only on social media, specifically Facebook and Twitter. Automated alt text is not good there. It's obvious it's not written by a human and it doesn't sound human." Furthermore, VIU1's comment stresses the need for suitability instead of automatically generated alt text; however, as discussed in the previous section, WCCs need to empathize with the web navigation experience of VIUs, but such empathy cannot be fostered when relying on automated approaches. Relatedly, (Gleason, Pavel, *et al.*, 2019; Gleason *et al.*, 2020) first experimented with a semi-automated approach for alt text suitability on Twitter (now known as "X") where automation only worked for memes, which were less hard to describe suitably in an automated manner, before advocating the use of crowdsourcing-based approaches in social media contexts.

Diving a bit deeper into the reasons for VIUs' low expectations, VIU2 highlighted the need for suitability, as alt text is often ignored: "What does "Click here" mean? We should actually know where this link is gonna take you and to have something in its label which indicates where you're going, because content using a screen reader is much more focused. I don't think I would be missing out on a great deal if alt text was all set to 0. I think I kind of ignore it most of the time." VIUs ignoring alt text due to its unsuitability, in fact, supports some WCCs' views (see aection 8.4.1.3) on the need for images to be marked as decorative to avoid Webpage navigation disruption that, at the same time, addresses a further barrier in alt text being left unlabeled, namely non-null alt text, with VIU2 stating that "The screen reader just ignores it, but if it's just been left unlabeled, I get unlabeled graphic, unlabeled graphic, unlabeled graphic—that's all the time!" Whilst the non-null alt text barrier is currently mentioned in certain accessibility resources (Caprette, 2025), again, it has not yet been formalized in academic literature. Another participant (VIU11) corroborates the burden of web navigation via screen readers being disrupted for the narration of alt text non-involving of any

functionality, stating that "The description needs to be functional. I'm not interested in the image being a scissors or a folder; I'd want to hear that it's a cut or a save. Or it can be decorative, so I mustn't listen to anything."

Suitability, therefore, becomes a graver concern when non-text content is also functional, e.g., an image that is also a web link; as per the comment above, if the alt text does not describe where the web link leads to, then the coin-flipping nature of web navigation is again evident. It is important to note at this stage that although WCAG highlights the need for the purpose of images to be described in alt text (W3C, 2023a), there is no suitable guideline on how to properly author alt text for images that are also web links despite available efforts (Gudhka, 2021), which are deemed inconsistent in different contexts. Another participant (VIU11) explains how this extends to alt text for graphs "For graphs, it should give you with one phrase the conclusion you draw from this graph and the information about where you can find the full-text description." This comment agrees well with WCCs' view (see section 8.5.1.2) on the need for alt text to link to a detailed plain text description of the data presented in the graph, rather than being more detailed itself. Importantly, the previously mentioned distinction between alt text and plain text with regards to the way those are treated by screen readers, is also highlighted by VIU10 "The problem with putting hugely detailed information into alt text is that for screen readers to browse that alt text line by line or word for word, you can't; you read it as a chunk." Importantly, it can be noted in this section too that sentiments can be tied to the profiles of the participants (Table 4); i.e., both VIU4 and VIU6 have more than five years of experience using screen readers and mention the unpredictable – coin flipping - nature of browsing the web via screen readers, as well as how it persists across experience levels.

8.5.2.2 Pseudo-experts squad (S-RQ1, S-RQ3)

Notably, the analysis revealed that the pseudo-experts squad theme is shared between WCCs and VIUs, which highlights how the latter put their trust in WCCs in terms of web accessibility. As VIU9 mentioned "Because I can't see, I have to trust that what you're telling me is exactly what is there," a comment which is in line with previous studies discussing how VIUs place a lot of trust in WCCs to have catered to accessibility, not least in relation to alt text (MacLeod *et al.*, 2017; Salisbury, Kamar and Morris, 2017). The above comment also stresses that a reason that such trust is forced upon VIUs is because they cannot know what is there to describe it in alt text. Interestingly, this contradicts past scholarly work about the need to renegotiate the role of VIUs and to turn them into alt text authors (Chisholm and Henry, 2005; Heylighen, Van der Linden and Van Steenwinkel, 2017), which aligns better with recent findings suggesting

that WCCs need to learn how to create more accessible websites as their preferred way towards a more inclusive web navigation experience (WebAIM, 2024b).

Similarly, the focus on WCCs engaging more with how to deliver accessible web navigation experiences is also highlighted in a comment by VIU7 who questioned the ability of assistive technology: "Does the screen reader have to make up for the mistakes that web developers are making? JAWS has tried to do that because they get a lot of feedback of their users and it's their job to try to improve that experience. Theoretically, assistive tech is not up to date to deal with all the accessibility errors, but I don't think that the screen reader is supposed to make up for that." Interestingly, this comment adds to previous claims in the literature that advancements in technology have outpaced advancements in assistive technologies (Stratton *et al.*, 2022), and again highlights that accessibility is neither the responsibility of VIUs, nor of assistive technologies. The "silver bullet" myth (Brewer, 2004), therefore, is not well-received by VIUs, which is consistent with recent findings that only a small percentage (14.1%) wanting advancements in screen readers (WebAIM, 2024a).

Another interesting point identified in the analysis was the need for equity with respect to web navigation, stressing in particular the limits of alt text and the need for suitability. Participant VIU5 highlighted that " … it [alt text] can never be as good as an image because it's a kind of translation in a way, it's a … so complex, but it should give you something because if it doesn't give you anything, better to do equal 0. They should never be done by AI, because as good as they are, only a person could identify and think, okay, how complex do I have to make it, what is the context, why do I need it, I think it's something that only a human being can do in this way." This comment further demonstrates that automatically generated alt text (e.g., by AI) is perceived as vastly inferior to manually authored alt text, which stresses the need for WCCs to train in suitable alt text authorship. In fact, past work corroborates that the suitability of alt text is very much dependent on context research (Miranda and Araujo, 2022; Muehlbradt and Kane, 2022), and WCCs are the only ones who can interact with said context to decide how it should be described as suitable alt text. A different participant (VIU11) in fact highlights the suitability gap in AI-generated and manually authored alt text, stating that "It's very important to know what is made by AI and what is real content. As in an extra label on alt text that tells me whether it's made by AI."

Section 8.5.1 highlighted that there is a mental resistance challenge among WCCs to dealing with alt text unavailability and unsuitability barriers. This is still relevant in the context of the present discussion, with VIU10 noting the need for WCCs to be trained in alt text suitability: "Education. First of all, the actual mechanics of writing alt text is easy. Adding the alt text is

easy. Getting the mindset that you want to actually add the alt text is the thing." Therefore, there seems to be a point among VIUs underlining that WCCs should be the only authors of alt text and appreciating that current efforts largely lean towards inaccessibility. Accordingly, these findings highlight that VIUs don't perceive themselves alone as adequate to be alt text authors, and further stress that this task is the responsibility of WCCs. We therefore argue that alt text authorship could benefit from more collaborative ways between VIUs and WCCs, where the former are positioned as evaluators and are able to reach out to WCCs about alt text barriers (see section 8.5.1.3).

Finally, when it comes to what kind of training WCCs should undertake to become "experts" in web accessibility, VIU11 (who is also a WCC) discussed that "The first thing when I want to write a good alt text is which information in an image description are important to what audience and then group them in a way that they are simple, understandable, language-wise, and solid, and that, objectively, will give me a good alt text, but of course the enemy of good is the better, but that is no concerns to us. What concerns us is that we've put a reasonable effort that leads us to a result above mediocrity and we have exercised all the correct guidelines for the authorship of a good alt text. The whether it could have been done better, well, everything could have been done better." This comment aligns well with this work's findings about the variability of such expertise among WCCs (see section 8.5.1), and points to the need to strive towards "pseudo-expertise" instead. The need for a common blueprint for WCCs that is more realistic and engaging than web accessibility guidelines is also highlighted in the above extract, and some indicative guidelines about alt text suitability are also provided.

8.5.2.3 Blindfolding (S-RQ2, S-RQ3)

Following on from the identified pressing need for training WCCs in web accessibility and the biggest challenge thereof being their mental resistance in so doing, the following comment from VIU3 highlights a mismatch between what WCCs include in alt text and how this is redundant for VIUs: "They think that I need to know the color, the length, the distance, or they say "People look like Colin Firth" … I've never seen Colin Firth!." This is particularly useful to pinpoint the need to foster empathy and understand web navigation via screen readers early on in the process, which interestingly, aligns with the findings for WCCs in Section 8.5.1.

The need to foster empathy is also highlighted by VIU2 who explained: "Whoever's deciding, you know, the developer, they need to think, as a screen reader user. Do they actually wanna know descriptions of all of these pictures? And I'll say the answer to that question is: "Probably not." There's always a certain judgment call to be made. Alt text equals 0 is a very

good starting point for all graphics, because it's giving some sympathy to the fact that I have to listen to all of this. That's the world I live in. It's a world which is audio and sympathy towards that is important and alt text equals 0 is a service, because you're saving me from all that stuff that I don't wanna listen to." This is indeed an important finding, as it highlights VIUs' preference to include null alt text, which it will indicate to assistive technology that an image can be safely ignored (W3C, 2016); its inclusion therefore can help avoid the interruption of VIUs' web navigation experience, as otherwise the screen reader would stop the navigation midway to narrate that an alt text is empty. Null alt text is in fact advised for decorative images (see section 8.5.2.1), but Lengua, Rubano and Vitali (2022) recently showed that distinguishing a non-decorative from a decorative image can be challenging for human authors and almost impossible for AI. In a related vein, participant VIU7 sheds light on another barrier relating to the misuse of alt text, not for decorative images but for images of text "My main problem is that people make images of text. It's not about if that image should have a text alternative; this image shouldn't have existed in the first place." This relates to the discussion on the difference between how screen readers treat alt text and plain text (see section 8.5.1.2), with the latter posing no navigation disruption barriers, and the identified barrier, namely image misuse, involves the ill use of images, which can present barriers as alternatives to text, which does not, and it is further corroborated by VIU10: "I'd want to read it, but I can't read it because it's not a text; it's an image." Similarly to the non-null alt text barrier, certain accessibility resources have mentioned the image misuse barrier (Bureau of Internet Accessibility, 2018), but again, to the best of the researcher's knowledge, this barrier is not yet formalised in the literature.

Staying with the web navigation experience of VIUs, participant VIU1 explained that such experience is fundamentally distinct from the visual experience: "They need to start looking at the image and describing it like a human who cares. Maybe if you used a screen reader you might figure it out. We need to educate people who don't use it or who don't know what it is. I don't know how you do that though … How do you educate people who don't know what it is?" web accessibility-related training should therefore first make this fundamental distinction clearer to foster empathy and, as such, address the afore-mentioned mental resistance challenge among WCCs. In an alt text context, for instance, VIUs mention how the language alt text is authored in results in contextual information being missed: "Language-wise alt text needs to be aligned to its surrounding context, e.g., in a site with comics and humor, alt text descriptions of images should equally have instances of humor" [VIU11]. Participant VIU7, in fact, stresses the key role of context in dictating how or if alt text should be authored beyond language

considerations: “It’s not the image that decides what is the text alternative, but for a big part it’s the context of which that image is used that influences whether you need the text alternative.” It is, however, important to note that alt text is not only accessed by blind people, and VIU3 emphasizes that it should be authored by taking into consideration all potential screen reader users: “Within the VI [visually impaired] community, there’s always a compromise: Enough to give me a hint and enough to give someone who is partially sighted or sight impaired sufficiency as well.”

Furthermore, the previously identified unhealthy foundation in terms of the motivation of WCCs to engage with web accessibility (see section 8.5.1.1) was also brought up with participant VIU7 stating: “I think there is some added value in don’t just … interpreting the guidelines. Sometimes you have to say that this is a failure according to the guidelines, but in reality no one really cares. That’s what we try to tell people do it not for compliance or for legal … If this motivates you then go ahead but I guess the best motivation is to have more customers and happy customers. Try to imagine the image is not there and what information do you lose, but it seems too analytic for people to do.” Training in this regard needs to be based on healthier benefits, such as happier users, and it should involve a way to get into the shoes of VIUs when experiencing web navigation to foster empathy.

Accordingly, the discussion so far indicates that WCCs are more suitable to author alt text, as unlike VIUs, they can see the non-text content that they need to describe: “Ask web designers: ‘Just close your eyes. Close your eyes!’ And you know that’s the picture of a banana, how would you tell yourself that’s a picture of a banana? Go backwards and then go forward, empty the alt text and let it describe it. You have an added advantage, because you can see it, but close your eyes and look at it from a blind person’s perspective for a second” [VIU9]. Effectively, this requires them to transition between the two web navigation experiences via a simulation of VIU experiences. This blindfolding simulation therefore is an essential part of any training for WCCs in web accessibility to motivate them towards accessibility in a healthier way, and further introduce them to guidelines about how to cater towards specific web accessibility barriers. Importantly, it is not implied that engagement with such a simulation will foster empathy; rather, the theme emphasises the need for accessibility solutions that use means to foster empathy before any training takes place, a finding that is in line with principles of the human-centered design process (Bennett and Rosner, 2019).

8.5.2.4 Trainability recommendations per VIUs

This section discussed the perceptions of VIUs in relation to web accessibility, not least in relation to screen readers and alt text. VIUs emphasise the resemblance of web navigation via screen readers with the flip of a coin, due to low effort allocation towards accessibility, resulting in VIUs having low expectations with regard to content being accessible via screen readers. Agreeing well with our WCCs (see section 8.5.1.4), VIUs also highlight a mental resistance as accessibility's worst enemy and further underline the need for WCCs to empathize with web navigation via screen readers as the first step to any accessibility-related training. Additionally, a few VIUs expressed a dislike towards the use of AI for catering to web accessibility, not least in relation to alt text, which they almost always consider unsuitable when authored by AI, while they also necessitate that WCCs are the only ones responsible for delivering accessibility, contradicting past evidence on assistive technology improvements and VIUs being actively involved in delivering accessibility that were considered as viable future avenues. Taken together, these findings point to the need for accessibility-related training that:

- Is the responsibility of WCCs or is a collaborative effort between WCCs and VIUs: Drive perceptions away from efforts that put the responsibility of ensuring the accessibility of uploaded web content away from the uploader (Section 8.5.2.2).
- Is initiated with a glimpse of web navigation via screen readers: Allows for empathizing with the nature of the experience of navigating the web via screen readers (Sections 8.5.2.1 and 8.5.2.3) in line with human-centred design principles.
- Is example-driven: Allows for understanding that accessibility cannot be perfect, but it should involve every reasonable effort in recognizing and staying away from inaccessibility in a plurality of contexts (Section 8.5.2.2).

8.5.3 Alt text suitability recommendations

Accordingly, a further set of recommendations for alt text suitability (Table 7) is proposed using codes from phase two of the reflexive thematic analysis process presented in Table 6, which are to the best of the researcher's knowledge, the first guidelines that compare and bring together the views of both WCCs and VIUs in the context of alt text suitability.

Table 7. Alt text suitability recommendations — web content creators ft. visually impaired users

Recommendation	WCC (# of participants)	VIUs (# of participants)	Example Extract
Context-specific	9	8	*It's not the image that decides what is the text alternative, but for a big part it's the context of which that image is used that influences whether you need the text alternative.* [VIU 7]
Decorative Case[a]	8	7	*Don't be afraid to mark things as decorative, you see far too much alt text on stuff that's decorative* and I think people are worried that they are gonna get it wrong [WCC 8].
Graph-specific	5	6	*For graphs, it should give you with one phrase the conclusion you draw from this graph and the information about where you can find the full-text description.* [VIU 11]
Functionality Prioritisation	6	5	*The description needs to be functional. I'm not interested in the image being a scissors or a folder; I'd want to hear that it's a cut or a save. Or it can be decorative, so I mustn't listen to anything*. [VIU 11]
Concise	4	5	*If you use too much information it might be necessary and it might be annoying to the people using screen readers as well, so the length of your alt text definitely needs to be considered; not too long but also not too short; that's not useful either.* [WCC 9]
Non-repetitive	3	4	*'An image of a cat,' because alt text describes an image, so you don't say: 'An image of,' you say: 'A cat.' It [the screen reader] knows it's an image or graphic, so you don't say: 'A photo of,' 'An image of;' you just say what it is.* [WCC 10]
Poetry Case[b]	3	4	*Decorative images enhance the appreciation of a webpage. Images of all kinds do. So, I think pretty much all images should have alt text. It's back to poetry, yeah?* [WCC 10]
Images ≠ Text[c]	0	4	*My main problem is that people make images of text. It's not about if that image should have a text alternative; this image shouldn't have existed in the first place.* [VIU 7]
Disability-specific	0	3	*Within the VI [visually impaired] community, there's always a compromise: Enough to give me a hint and enough to give someone who is partially sighted or sight impaired sufficiency as well.* [VIU 3]

Recommendation	WCC (# of participants)	VIUs (# of participants)	Example Extract
Author Transparency	0	2	*It's very important to know what is made by AI and what is real content. As in an extra label on alt text that tells me whether it's made by AI.* [VIU 11]
Plain Language	2	0	*They need concepts broken down into plain language; they shouldn't be reading like all these things being written on a graduate level* [WCC 9]

[a] Null alt for decorative images. [b] All images should have alt text, no image is decorative. [c] Communicate to the author that the image should be converted to text, as images of text should not exist.

The above table is revealing in several ways. First, it contradicts past evidence (e.g., Harris, 2020; Hanley *et al.*, 2021) on the persistence of a mismatch between the perceptions of WCCs and VIUs on alt text suitability, as this mismatch is very negligible and only evident for less regularly reported guidelines in the table. However, it is important to heed that the sample of WCCs in this work had an average of nine years of experience in web accessibility, and it is thus not as surprising that their views align well with those of VIUs. It can, thus, be surmised that a mismatch persists when expertise in web accessibility has been gained through an unhealthy foundation, as discussed in Section 8.5.2, about the low encounter rate of suitable alt text and the coin flipping nature of web navigation via screen readers overall. Second, it highlights website context as the main determining factor of how and if alt text should be authored followed by the need to mark images as decorative if there is no functionality, which also address the afore-identified non-null alt text barrier (Section 8.5.2.1). Similarly, the need to replace images of text with plain text, deferring to the image misuse barrier (see Section 8.5.2.3), is stressed by VIUs, while the need for alt text to be concise and non-repetitive of surrounding content is stressed by both groups. Finally, both groups highlight that alt text for graphs should be treated uniquely via a brief description that includes the type of the graph and any conclusion that can be drawn from it, and information about where a detailed description in plain text can be found.

8.6 Chapter summary

This chapter presented a qualitative user study with 11 WCCs and 11 VIUs that was conducted as part of this thesis to address the reported mismatch in relevant literature between their views when it comes to web inaccessibility overall, and then more specifically, in relation to alt text suitability. A reflexive thematic analysis approach was followed (Table 6) presenting an

analytic narrative anchored to theory and the S-RQs of the study. The findings stressed that the mismatch between WCCs and VIUs stems from the formers' lack of experiential understanding of web navigation via screen readers. Both groups agreed that alt text barriers are most often due to no effort being focused on accessibility rather than task complexity, e.g., it is not difficult for a sighted WCC to tell whether an image only depicts text. Therefore, it is not surprising that both participant groups suggest increased efforts to gain pseudo-expertise in accessibility (Sections 8.5.1.3 and 8.5.2.2) rather than less efforts aimed at accessibility expertise, especially considering the variability of such expertise. This highlights the need for the proposed GWAP solution, as an approach for recruiting non-experts via crowdsourcing and training them under a common blueprint (i.e., context definition (see section 4.3.1)). Training will be example-driven as per the trainability recommendation presented in this study (Sections 8.5.1.4 and 8.5.2.4). Accordingly, Chapter 9 discusses the proposed GWAP solution that was developed based on the literature findings, as well as the findings deriving from this user study.

Chapter 9. System design and framework

9.1 Introduction

The landscape of the literature around alt text barriers presented in the previous chapters has shown that despite widespread acknowledgment of missing alt text as a significant challenge, the presence of unsuitable alt text remains an equally pervasive yet often overlooked issue. Hence, the need for the alt text suitability recommendations presented in the previous chapter. The recommendations were in fact in line with literature findings on the concept of 'context in which the image is used in' being central to alt text suitability; however, it was also shown that web accessibility guidelines are ambiguous with regard to suitability (McCall and Chagnon, 2022), and there is a general acceptance that authoring suitable alt text is a complex task (Hanley *et al.*, 2021). Whilst a plurality of solutions to alt text barriers have been proposed in the literature, ranging from manual to automated and crowdsourcing approaches, the latter appear most promising to engage non-expert alt text authors, as expertise in alt text suitability is ambiguous. This is due to such approaches' inherent ability to leverage the collective efforts of a large and diverse number of participants to enable the generation and evaluation of alt text at a much faster rate than manual efforts, whilst ensuring that alt text is more representative of different views leading to more contextually appropriate and meaningful descriptions.

Additionally, the inherent difficulty of authoring alt text suitably prohibits the outright handing of the task to non-experts without prior training. To address these challenges, GWAPs have emerged as a promising approach leveraging the benefits of crowdsourcing (see chapter 5). GWAPs excel at handing out crowdsourcing tasks to non-experts, as they excel training users (Tuite, 2014), which aligns with the need to turn users into pseudo-experts (see previous section) and has yielded promising results for complex tasks (Aliady *et al.*, 2022). Additionally, and unlike other crowdsourcing approaches that may rely on monetary incentives or task-based participation, GWAPs foster intrinsic motivation through gameplay enjoyment, which sustains consistent user engagement over time. These systems can therefore produce diverse, high-quality, and more contextually appropriate alt text at scale. Proposing a GWAP-based solution therefore appears well-suited to address the reported dual challenge of unsuitable alt text and inadequate authoring practices. Accordingly, a novel GWAP solution – TagALTlong – aiming to bridge these gaps by showcasing its effectiveness in generating reliable and context-driven alt text descriptions at scale is proposed in this chapter.

Through this lens, the work in this chapter underscores the importance of integrating collaborative approaches to ensure human-centred alt text descriptions. Whilst there have been past similar efforts with noted implications for improving the suitability of alt text (see section

5.2.2), these focused on different annotation tasks, such as image feature and virtual world object metadata annotation, and context was not considered. As such and to the best of the researcher's knowledge, there is no recent GWAP for context-driven alt text annotation.

Accordingly, this chapter discusses the infrastructures comprising the system, underpinning the implemented solution (TagALTlong) presented in this thesis. The chapter first presents the detailed design of TagALTlong, which has been designed with an aim to efficiently generate suitable alt text annotations through a context and community-driven approach utilising the definition discussed in Section 4.3.1. The cloud infrastructure that was used to host the backend of the game, as well as the database, is presented next. Finally, it was necessary that a machine learning (ML) infrastructure was included, due to the discussed need for automation in the case of alt text generation to address the reluctance for authoring alt text and the need to scale its generation on par with the increase of multimedia content on the Web (see section 4.5). The ML infrastructure therefore presents the proposed ML models (incl. architecture and pipeline) (see section 9.2.3).

9.2 Framework architecture components

As explained in Chapter 6, it was necessary that TagALTlong's output be used for the training of an ML model with the aim of approximating the automation of human-centred alt text. The previous findings from the literature and the user study (see chapter 8) pointed towards several requirements that need to be covered by the proposed system solution.

Table 8. Mapping table of system requirements to architecture components

System requirements	Related findings	Architecture components
GWAP interface for large-scale data collection with pseudo-expert users	Expert opinions vary; Alt text suitability is complex; Large-scale datasets are needed to train V2L models (Section 4.4)	Frontend services: Unity C#; GWAP for human-curated alt text generation
Context-driven training data	Lack of context in SoTA training datasets (Section 6.3); Context is central to alt text suitability (Secton 4.1)	Dataset A: context prompts; Dataset B: context
Storage of training data, supporting peer review and large-scale data collection	GWAP design typology (Pe-Than, Goh and Lee, 2015), and framework (Bellscheidt *et al.*, 2023) (Section 5.3.1); Trainability recommendations (Chapter 8)	Backend services: Oracle Cloud Infrastructure (OCI); Apache Web server (PHP backend); MySQL database
ML pipeline for automated alt text generation	Alt text author reluctance; Multimedia content increase on the Web (Section 4.5); AI surge	Image feature extractor; ML model for automated alt text generation

The architecture components in the above table related to the components of the architecture of the framework of the system (see Fig. 10 below), where the three infrastructures, i.e., GWAP, cloud and ML infrastructures, are shown.

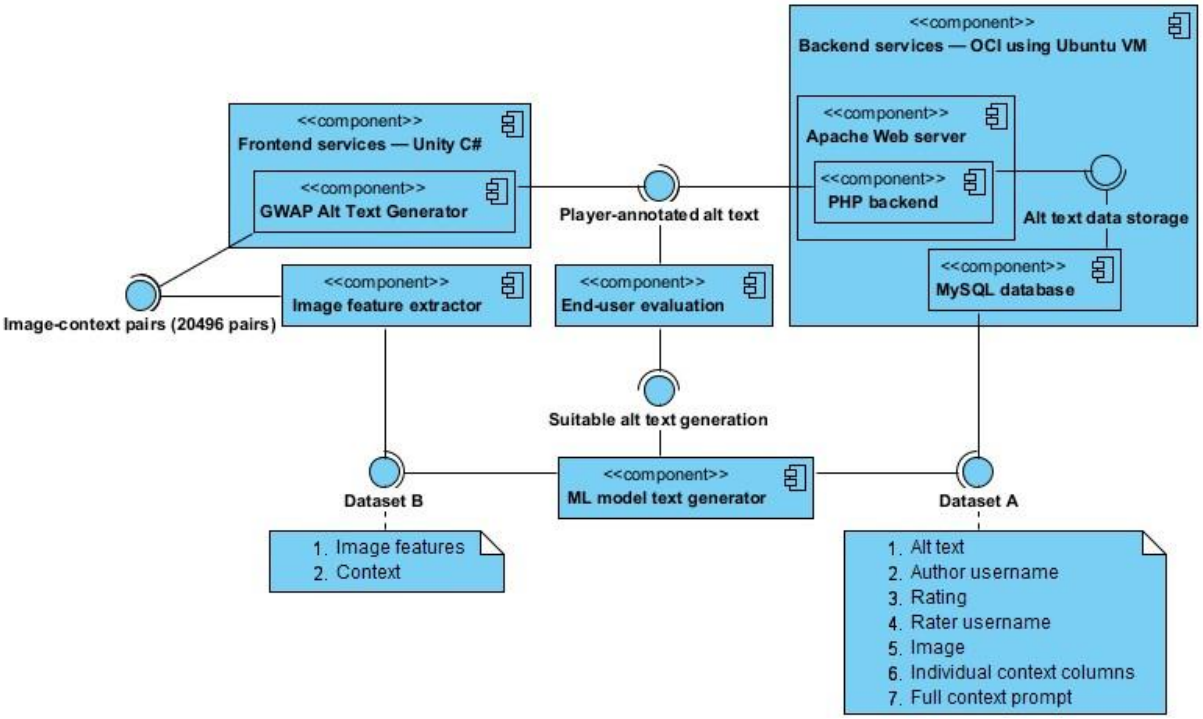


Fig. 10. UML component diagram of the framework architecture and the inter-connected infrastructures

Deferring to the above diagram, the following sections in this chapter explain the role of each infrastructure in the context of the framework.

9.2.1 GWAP infrastructure

This section presents the detailed design of TagALTlong: the GWAP developed in this thesis.

9.2.1.1 Game design framework

The proposed game (TagALTlong) belongs to the GWAP paradigm, thus largely draws on Pe-Than, Goh and Lee (2015)'s typology, hence the design approach followed is focusing on three dimensions, i.e., gameplay mode, gameplay structure, and data. Accordingly, TagALTlong's gameplay is asynchronous, which means that players do not need to be online at the same time to play. Interaction between players is anonymous and indirect, which is reflected in the game by asking players to rate alt text authored by other players without knowing the identity of the author. The structure has been designed to be a standalone singleplayer game, but through a collaborative data collection approach wherein players can author alt text on their own but can also rate others' previously authored alt text, and vice versa. The collection of data has been

designed to be open-ended whilst a redundancy mechanism (i.e., assigning the same task to many players until consensus is observed in their outputs) has been implemented for data control. The game also utilises multilevel peer review through the alt text rater mode as a verification mechanism, while iterative improvement is used for the incremental refinement of player output via the collection of a plurality of annotations for the same combinations of image-context-alt text. These are summarised as the framework in Table 9.

Table 9. Mapping the game design to Pe-Than, Goh and Lee (2015)'s typology

Category	Dimension	TagALTlong Approach
Gameplay Design	Mode	**Asynchronous** – Players do not need to be online at the same time
	Interaction Type	**Anonymous & Indirect** – Players rate alt text written by others without knowing the author's identity.
	Structure	**Standalone Single-Player** – Players play individually but contribute collaboratively through data annotation.
Collaboration Mechanism	Player Contribution	**Peer Review & Consensus** – Players rate and refine alt text written by others, promoting collective improvement.
Data	Collection Approach	**Open-Ended** – Players generate freeform alt text using context definition (section 4.3.1)
	Redundancy Control	**Redundancy Mechanism** – Multiple players complete the same task to ensure consistency in annotations.
	Verification	**Multilevel Peer Review** – The "alt text rater" mode acts as a quality control and verification process.
	Data Refinement	**Iterative Improvement** – Multiple annotations for the same image-context pair help refine the final alt text.

The above design choices are in line with past evidence on players in GWAP contexts preferring to 'wing' how to play correctly (Games and NLP, 2020), the need to minimise cognitive effort in such contexts (Bellscheidt *et al.*, 2023), the lack of motivation to engage with such training in non-gameful contexts due to 'task dullness' (Hanley *et al.*, 2021; Mack *et al.*, 2021), and the lack of consensus on what constitutes suitable alt text (Petrie *et al.*, 2011).

9.2.1.2 Environment and game flow

The game is structured to guide users through a clear and consistent interaction flow, beginning with a login/registration process (Fig. 11a) where they are first asked to read the online Participant Information Sheet (PIS) and the Consent form before consenting to take part in the study and to consequently register an account. Upon successful registration/authentication, players access a main menu where they select their user role, choosing to participate either as a content producer (Alt text author) or a content analyser (Alt text rater) (Fig. 11b). First-time users are encouraged to complete a brief onboarding tutorial (Fig. 11.c), which introduces the primary goal of alt text (Bellscheidt *et al.*, 2023), instructions about writing appropriate alt text

(Silktide, 2023; Cionca and Kohler, 2024; W3C, 2024a), guidelines about context based on the proposed definition (Fig. 11.d), as well as the gameplay mechanics for both roles.

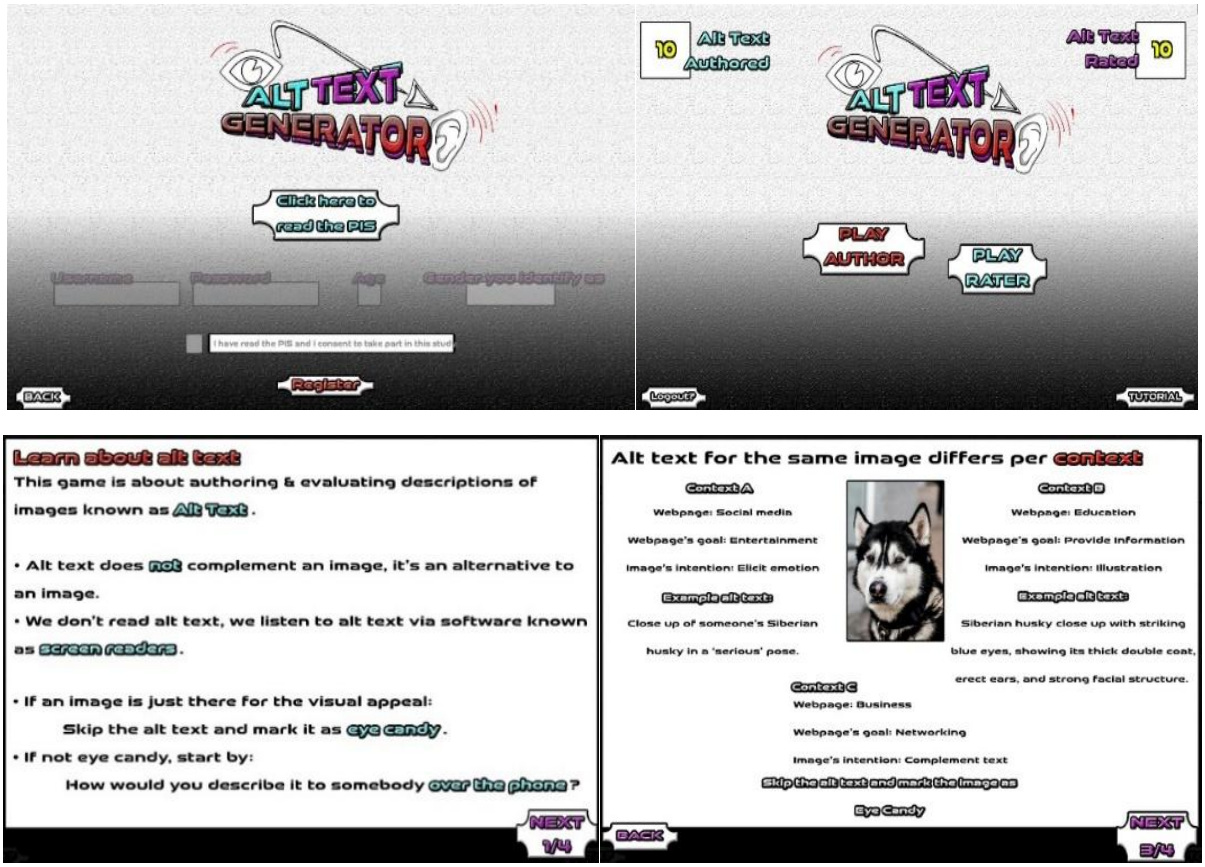


Fig. 11. Key game pages and popups. (a) Registration/login page (top left); (b) Main menu page (top right); (c) First popup of the tutorial (bottom left); (d) Context-related tutorial (bottom right)

As an Alt text author, players are presented with a random combination of an image and a context description derived from the structured context model proposed earlier (Section 4.3.1), and are prompted to write a suitable alt text description based on the image and its contextual framing, or mark it as 'Eye Candy' if they consider the image to be decorative in that given context (Fig. 12a). Upon finishing, players are directed to the Alt text rater role task, where as a rater they are shown random combinations of images and associated alt text descriptions authored by other players along with the relevant context, and are asked to rate their suitability using a predefined rating scale (1-5 stars) (Fig. 12b). They finally can submit their ratings. Raters are not shown the identity of the authors, ensuring anonymous, unbiased peer assessment. This setup enables asynchronous and collaborative data generation and assessment, contributing to the creation of a rich and context-sensitive dataset of alt text annotations.

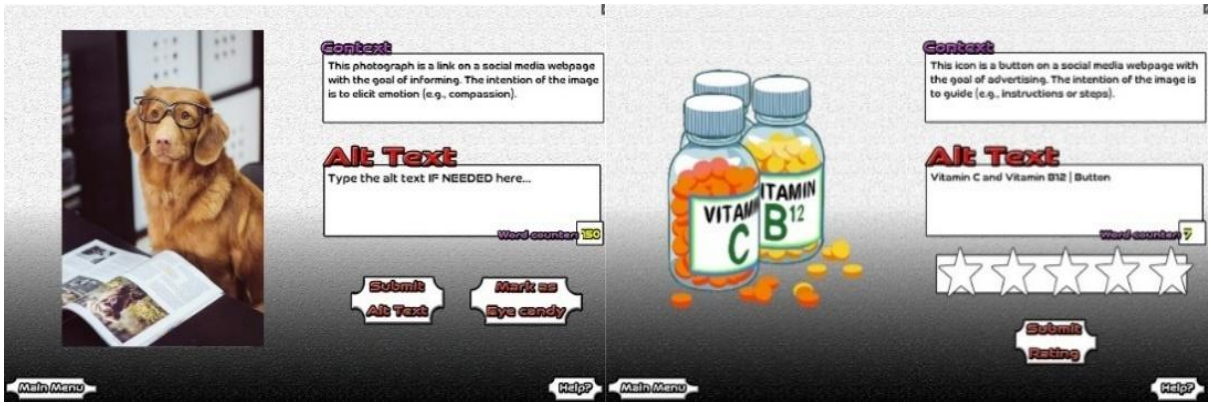


Fig. 12. Author and Rater tasks. (a) Authoring alt text description (left) (b) Evaluating alt text description (right)

9.2.1.3 Context generation approach

The importance of a structured context definition in tandem with the need to incorporate its semantic adaptation into solutions to alt text barriers was highlighted both in the literature (Chapter 4) and the findings from the first user study (Chapter 8). Accordingly, to support the creation of suitable and context-driven alt text annotations, the game employs a structured context generation approach that simulates real-world usage scenarios. In practice, each image presented to players is accompanied by a randomly generated context prompt, composed of the five interrelated factors in the proposed context definition 'altC' in Section 4.3.1 and various illustrating values per factor (Table 9), which reflects how images appear in diverse digital environments. These values were informed by prior work on the types of images and webpages that screen reader users expect to find and interact with alt text descriptions (e.g., Morris et al. (2016); Stangl et al. (2018)).

Table 10. Mapping the altC contextual factors to in-game context prompt values

Contextual Factor	Values
Image Type	Photograph, Graph, Logo, Painting, Icon
Function	Decorative (None), Functional (Link, Button)
Webpage Topic	Social Media, Education, Health, Business, Sports, Travel
Webpage Purpose	Informing, Learning, Entertaining, Selling, Advertising, Networking
Image Intent	Complement (e.g., enhance text context), Illustrate (e.g., show a product), Guide (e.g., instructions), Elicit emotion (e.g., compassion)

These factors and values are assembled into a **natural language prompt** in the game based on the following structured template:

This [*Image Type*]
is a/is found[13] [*Function*]
on [*Webpage Topic*] webpage
with the goal of [*Webpage Purpose*].
The intention of the image is to [*Image Intent*].

This dynamic generation of contextual scenarios ensures diversity in gameplay and helps collect alt text annotations that align with real-world web environments. To illustrate, using the aforementioned prompt template and Table 9, the game would generate the following context prompt: "This photograph is a link on a social media webpage with the goal of informing. The intention of the image is to elicit emotion" as a potential combination of more context-rich descriptions. Finally, on the other end, the resulting alt text annotations are fed into a feedback loop, which is effectively achieved through an uncapped data curation process (no limit to how many players rate the same image-context combination), as the proposed approach aims for curation through the collection of diverse opinions, aligning well with the variability of expertise on alt text suitability. A detailed account of the technicalities of the implementation is included in Appendix C (GWAP backend) and Appendix D (GWAP frontend).

9.2.2 Cloud infrastructure

Deferring to the UML component diagram revealing the framework of the solution proposed in this thesis (see Fig. 10), the backend of the GWAP was implemented using PHP and was hosted on a Virtual Machine (VM) on Oracle Cloud Infrastructure (OCI), with SQL server running on the same VM for managing the game's database. This architecture enabled a collaborative, scalable, and efficient environment for alt text generation and annotation. In fact, the choice for the game to be embedded in a webpage necessitated the use of a cloud-based infrastructure to host the backend and database for the game; thus, OCI was used to help realise this, as it offers a free subscription affording a remote server, in this case, an *Ubuntu* VM, with enough memory space for the purposes of the game. Several steps were, however, needed before this was made possible, which are included in Appendix E for completeness.

[13] if decorative (i.e., not functional), then 'is found' is automatically used in the prompt

9.2.3 Machine learning (ML) infrastructure

The final component of the framework involves an ML infrastructure, which involves two AI models:

- The first model — **HumanALT-O-matic** — aims to learn to automatically generate alt text comparable to average human-level quality based on the data generated by the players through the GWAP (Section 9.2.3.1).
- The second model — **ContextALT-O-matic** — aims to learn to automatically generate more context-driven alt text descriptions based on the GWAP-generated dataset and the use of context prompts (Section 9.2.3.2).

This section presents an overview of the architecture and training of the AI models. It should be noted that the contributions of this work and thesis are **not** to the ML field, but existing ML tools were used to produce a state-of-the-art approach for automatically generating alt text that is close to a human average. The originality of the ML infrastructure is thus in demonstrating the value of the GWAP-generated dataset to improve the automated generation of human-level and context-driven quality alt text through the training of the models discussed in the following sections. This is to validate the human-centred approach to data curation (Chapter 10), rather than web-scraped datasets, which are prohibitive for accessibility (see chapter 6). The need for an ML infrastructure is further motivated by the need to address the reluctance of WCCs to author alt text (Williams *et al.*, 2022) by automating the approach. Additionally, since no prior work has trained V2L models on GWAP-generated, context-driven alt text data (Chapter 6), the models presented in this work are a proof-of-concept to address this gap. Thus, whilst TagALTlong is a standalone crowdsourcing solution to alt text barriers, the ML infrastructure introduces a proof-of-concept for automating suitable alt text generation through AI based on a dataset that is both human-curated and context-driven.

9.2.3.1 The HumanALT-O-matic model: Architecture and training

The data generated by players via TagALTlong, as part of the user study described in the next chapter, comprised images, context prompts, alt text descriptions and rating scores, which were used at different stages of the pipeline.

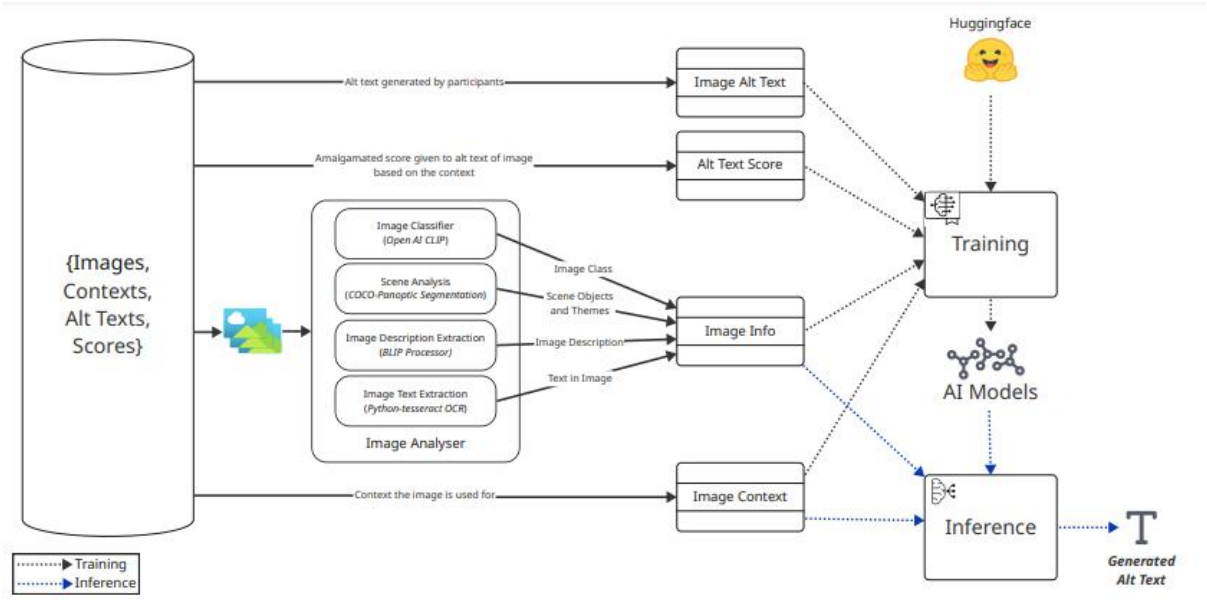


Fig. 13. HumanALT-O-matic pipeline overview

As shown in the above figure, the images were processed by an image analyser, which extracted image information (i.e., image class, scene objects and themes, image description, text in the image) that were then used in tandem with contextual information to generate alt text. Within the image analyser, Open AI CLIP was used for assigning a class to the image, complemented by COCO Panoptic Segmentation to identify thematic elements in the scene, as well as a BLIP Processor and a Python-tesseract OCR for generating an image caption and identifying text in the image, respectively. This image information and their respective context prompts were used to train a T5-small model. Two versions of this model were used, i.e., an initial **control** version of the model that was off-the-shelf and was used without training, and a second **trained** version that was trained and fine-tuned on the GWAP-generated dataset. Therefore, the data deriving from TagALTlong gameplay (player-authored alt text descriptions and players' amalgamated rating scores for image-context-alt text tuples) was used in tandem with the image information generated by the image analyser and the context prompts to train the T5-small model. It was thus made possible to compare the output of the control and trained versions of the model to assess whether the use of a GWAP-generated dataset to train the model is an effective approach to approximate average human-level quality alt text. This was assessed in the evaluation of the performance of the models (Chapter 11).

To support the automated generation of context-driven alt text, this two-model approach used a modular two-stage architecture (see Fig. 14 below). In the first stage, a BERT-based model is used to classify each image as a decorative (eye candy) or a non-decorative image, routing it to a different alt text generation pathway based on this classification. Non-decorative

images proceed to the second stage, where the T5-small model generates alt text using a distinct approach for each version of the model: the control version uses only image information and context prompts, while the trained version also uses player-authored alt text and ratings.

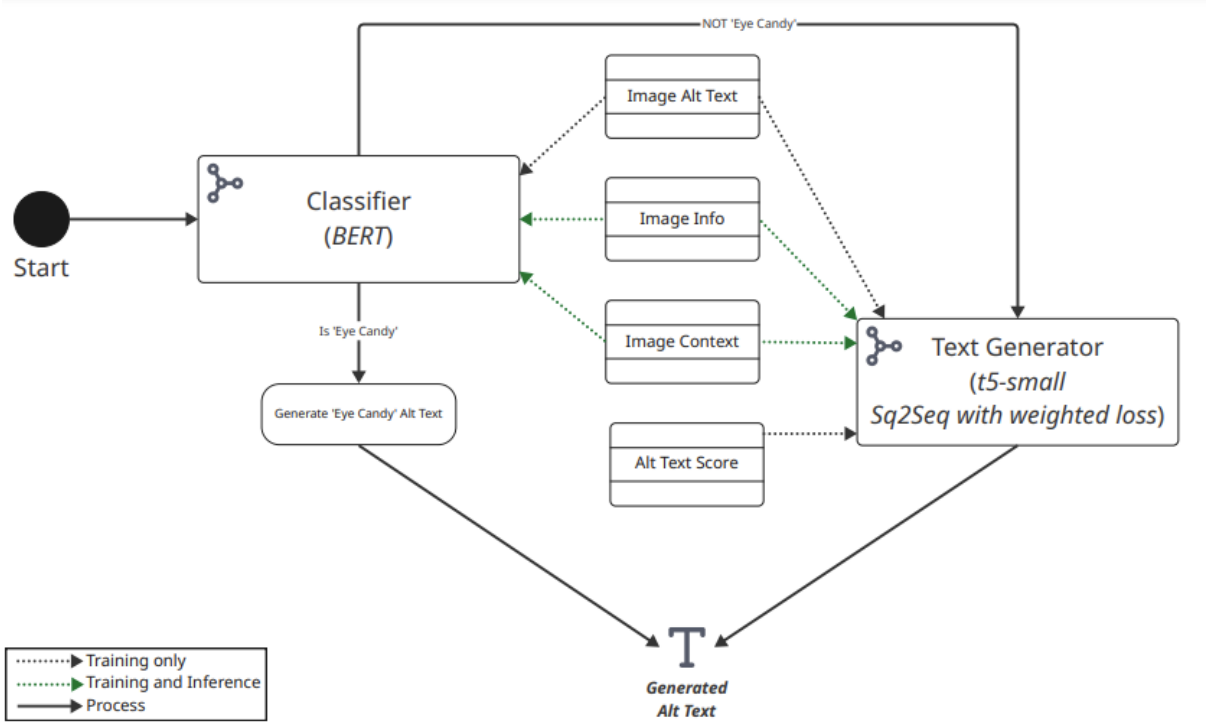


Fig. 14. HumanALT-O-matic architecture

The above figure shows how the BERT-based classifier distinguishes between decorative and non-decorative images, which is, to the best of the researcher's knowledge, the first effort to attempt this distinction for automated alt text generation. This distinction is crucial and in accordance with web accessibility literature and the findings from the first user study (see chapter 8), and it is further highlighted in an automated context due to the reluctance of web content creators to author alt text. For non-decorative images, the T5-small model uses a sequence-to-sequence (Seq2Seq) architecture with a weighted loss function to generate alt text. This translates the input of image information, context prompts, and GWAP-generated data (player-authored alt text and rating scores) into automatically generated alt text. The weighted loss function further treats alt text descriptions with higher player rating scores with higher importance. This feature was aimed at closely capturing the aggregated rating scores from the GWAP, translating them into a blueprint for the model to generate alt text descriptions close to average human-level quality. Importantly, 30% of the dataset was used to train the T5-small Seq2Seq model, while 10% of this (approx. 3-4% of the whole dataset) was used for validation.

Then, in inference, the model was applied to the entire dataset to generate an alt text description for all image-context pairs in the cleaned dataset.

9.2.3.2 The ContextALT-O-matic model: Architecture and training

Whereas HumanALT-O-matic leveraged human ratings to approximate average human-level quality, this second model was designed to test the role of context itself by training it to consider context during alt text generation. Two otherwise identical variants were trained that differed only in whether the training data contained contextual information. For this, an image-as-tokens V2L model, which can natively parse images as well as textual data in its token stream, was fine-tuned on the GWAP-derived preference dataset via contrastive preference learning (Direct Preference Optimization, DPO). The training objective was to prefer alt texts with higher human-given rating scores over those with poorer ratings.

The model selected for training, Qwen2.5-VL-3B-Instruct (Bai *et al.*, 2025), already produced alt text–like outputs before fine-tuning. This makes the GWAP preference data well-matched for post-training: it shapes generation toward human-preferred alt texts while leveraging the base model's ability to read rich natural-language context. In other words, the model primarily needs to learn how to apply context correctly rather than to parse it. The GWAP-generated dataset is suitable for reward-modeling techniques such as Reinforcement Learning for Human Feedback (RLHF) and for contrastive preference objectives such as DPO (Rafailov *et al.*, 2023), both of which steer models toward human-preferred outputs. DPO was adopted due to its simplicity, its stability on small- to medium-sized datasets, and its ability to train without a separate reward model, making it well-suited for the narrow task domain of alt text generation. Fig. 15 below shows the architecture of ContextALT-O-matic.

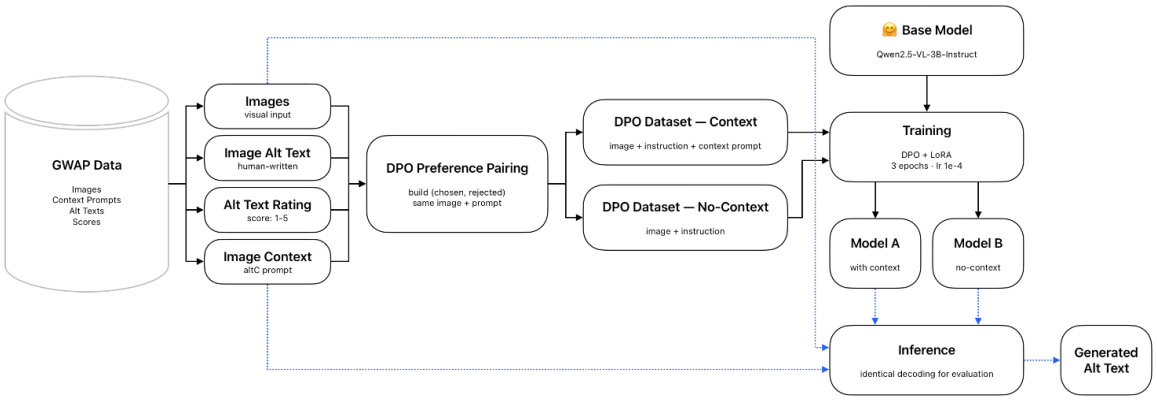


Fig. 15. ContextALT-O-matic architecture

The DPO dataset consists of pairs of image–context–alt text tuples: one with a higher human-given rating score and one with a lower rating score. Image and context are held constant within each pair, letting the model learn which differences in alt text lead to higher ratings. To examine whether training on context prompts aids context-aware alt text generation, two otherwise identical models were trained on the same DPO recipe, differing only in the exclusion of context prompts from the latter, which resulted in a context-aware and a no-context model. Both were trained with Low-Rank Adaptation (LoRA) for three epochs, at a learning rate of 1e-4. The goal of this setup was to evaluate whether context-aware training helps the model better utilize contextual information, rather than solely relying on its inherent understanding. While the base model can make use of contextual information without task-specific fine-tuning, subpar performance of the base model could improve through contrastive learning. Conversely, if the base model can already fully utilise context prompts, significant improvements are unlikely. To better understand the role of context during training, the altC context-presence was measured through binary classification on the outputs of the context-aware and no-context models. To accommodate dataset sparsity, 90% of the GWAP-generated data were used for training and 10% held out for validation. The context-presence evaluation dataset was generated from a random subsample of image-context pairs that received no human-authored alt text in the TagALTlong GWAP.

9.3 Chapter summary

This chapter presented the design of the system and related framework proposed in this thesis to address alt text barriers. The framework comprises three components, i.e., a GWAP, a cloud and an ML infrastructure, which work together for the automated generation of suitable alt text. Suitable alt text in this context refers to automatically generated alt text descriptions that are

close to average human-level quality. First, a user study will be conducted where players will be recruited to play the GWAP developed in this work to collect player-authored alt text and rating scores (see chapter 10). The backend of the GWAP, as well as the database where the output generated by the players is stored, are hosted in a remote cloud server. This output is used to train an AI model to automatically generate alt text descriptions, and its potential to use the GWAP-generated dataset to approximate average human-level quality will be assessed through a user study, where the output of the trained model will be compared with a control version of the same model that generated alt text based on pure image processing (see chapter 11). Accordingly, the following chapters present the design and findings of the user studies for evaluating the effectiveness of the GWAP in generating context-driven alt text (Chapter 10) and the ability of the trained model to outperform pure image processing in quality of generated alt text descriptions (Chapter 11).

Chapter 10. Second user study: Evaluation of alt text annotations through a context-driven game-based approach

10.1 Introduction

The previous chapter presented the workflow between the three components of the solution proposed in this thesis to address alt text barriers. Therefore, this chapter presents the user study that was conducted to validate the effectiveness of the implemented solution in generating context-driven alt text descriptions at scale (RQ3). The primary aim of the study was to assess the quality and diversity of the alt text annotations generated in relation to the context definition (Section 4.3.1), and to gather insights into the usefulness and perceived value of the associated context prompts. The study addressed following sub-research questions (S-RQs):

- S-RQ4. How effective is the implemented solution in generating alt text descriptions at scale?
- S-RQ5. How does the use of structured context prompts influence the quality of player-authored alt text?
- S-RQ6. How does the descriptive verbosity of player-authored alt text influence its quality perception?
- S-RQ7. What levels of consistency or divergence emerge among player ratings?

Data from the study included annotation content (authored alt text) and rating distributions. The study design, data collection, and analysis processes are detailed in the following sections.

10.2 Participants

Participants were recruited through convenience sampling, primarily via academic mailing lists, social media posts, and direct invitations within the researcher's institutional networks. In addition, snowball sampling was used to increase the number of participants. Participation was voluntary and no financial incentives were offered. Interested participants were prompted to contact the researcher via email to receive preliminary instructions and the web link to play the game. A total of N = 125 unique participants were recruited from 23 January to 5 March 2025 aged between 18 and 77 (M=30, SD=12.72) years old, and included 76 participants identified as male, 44 as female, 1 as other and 4 preferred not to say. For clarity, gender was used in this research in accordance with the definition of the World Health Organization (WHO) (2025).

10.3 Study design and procedure

Ethics approval was granted by the researcher's institutional Research Ethics Committee (Ref: 41665-A-Feb/2025- 53701-1) (see Appendix G). The study adopted a quantitative, exploratory

design to evaluate the feasibility of the implemented context-driven GWAP for alt text annotation. The objective was to assess the quality and consistency of alt text descriptions and peer ratings collected via through gameplay in a naturalistic, unsupervised setting allowing participants to access and play the game in their own time and using their own devices. No manipulation or experimental intervention was introduced; instead, the evaluation was exclusively based on the data captured through the game's backend during gameplay sessions. The design enabled large-scale, asynchronous participation and supported data validation through redundancy and multi-level peer review mechanisms embedded in the game structure. All data were generated exclusively through in-game interactions.

Accordingly, participants were provided with a private link to access the game on the itch.io platform. Upon visiting the game, they were first prompted to read the Participant Information Sheet (PIS) and Consent form, and following that, they were able to register an account. They were then asked to provide electronic Consent by agreeing to take part in the study. Only participants who provided consent were able to register and participate in the study. Following the game flow presented in Section 9.2.1.2, if this was a participant's first time playthrough, the game directed them to a tutorial to familiarise them with the mechanics and objectives of both Author and Rater modes. It provided clear instructions on how to write alt text descriptions, how to rate other players' alt text, and information about the meaning of context in this work. First-time participants were required to go through this tutorial before they could start playing. They could take as much time as needed to complete the tutorial and could revisit it if necessary, through the game menu (also as returning players).

After completing the tutorial, first-time participants were directed to Author mode where they were shown a randomly selected image paired with a pre-generated context prompt, as described in Section 9.2.1.3. Participants were tasked with composing 10 suitable alt text descriptions for image-context pairs before proceeding to Rater mode. Following that, in Rater mode participants were presented with alt text descriptions generated by other players which were paired with corresponding images and context. Participants were then asked to rate the suitability of 10 alt text descriptions based on a five-point Likert scale. Once 10 alt text descriptions/ratings were submitted, participants could continue to the next image or switch modes. For returning participants, the game allowed them to access either Author or Rater mode based on their previous gameplay history, enabling them to continue contributing without the mandatory sequence. Upon submission, all inputs (alt text submissions and ratings) were recorded in the backend database. It is estimated that each gameplay session lasted approximately 20-30 minutes depending on the participant.

10.4 Results

The various levels of analysis are presented in this section, beginning with an analysis of the effectiveness of the implemented solution (S-RQ4) followed by an analysis of the influence of context prompts (S-RQ5) and the descriptive verbosity of alt text descriptions (S-RQ6) on the quality perception of player-authored alt text; and finally, the level of consistency among player ratings (S-RQ7)

10.4.1 Alt text generation effectiveness (S-RQ4)

First, an aggregated overview of collected data is presented in this section prior to discussing the subsequent analysis and findings. The study participants authored 1208 and rated 1836 alt text descriptions, respectively, over 6 weeks. One hundred and fourteen successfully completed the user study requirements (authoring and rating at least 10 alt text descriptions) showing a 91% completion rate. The distribution of the number of rating scores for each alt text description varied with 95 descriptions receiving as many as five to thirteen rating scores, while 728 alt text descriptions received at least one rating score. Each alt text description received 2.52 ratings on average (SD = 1.92). Fig. 16 below is a cumulative overview showing the weekly growth of TagALTlong's output in player contributions over the six-week period.

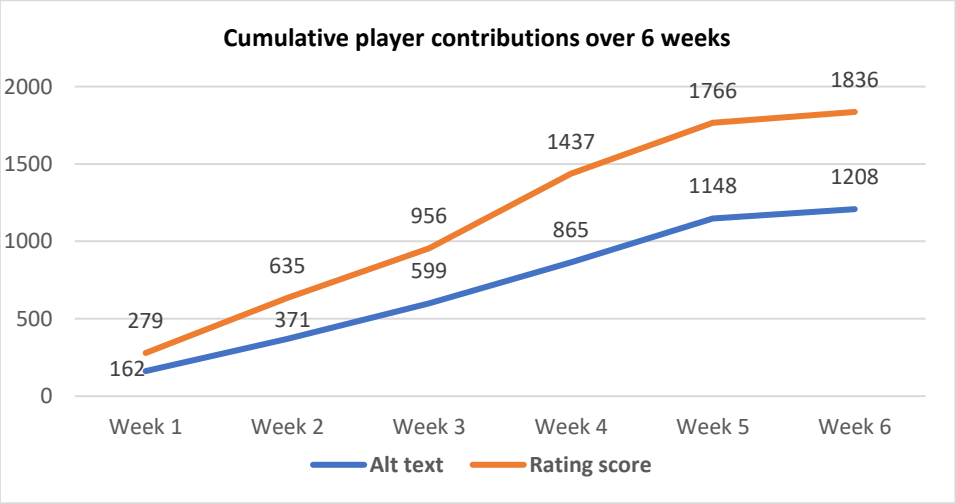


Fig. 16. Weekly growth of player-annotated alt text descriptions and rating scores in TagALTlong over 6 weeks

Accordingly, player contributions in alt text descriptions and rating scores show an overall increasing weekly trend over the six-week period, which was most notable during the first five weeks. Over this period, average weekly player contributions were 201 alt text descriptions and 306 rating scores, which were supported by an average weekly recruitment of 21 players. Coverage-wise, 31 out of 32 images received at least one alt text description and one rating

score, while 28 and 29 images received at least five alt text and five ratings, respectively. On average, each image received 39 alt text descriptions and 59 rating scores, while 28 out of 32 images received both at least five descriptions and at least five scores, which, if taken together with the above numbers, highlight the potential of the approach to scale alt text annotation based on these promising early results.

10.4.2 Data cleaning

Prior to further analysis to address the remaining research questions, a data cleaning stage was deemed as a necessary step for the rating population of alt text descriptions that received at least two rating scores (n=430 descriptions and 1538 rating scores). The analysis was limited to alt text descriptions that received at least two independent ratings to ensure greater reliability and robustness in the aggregated scores. Descriptions rated only once do not provide any measure of agreement or variability, which are crucial to identify outliers, gauging confidence in user perceptions, and performing meaningful statistical comparisons. Thus, and in line with recent practice to subjective assessment of disagreement among crowdworkers (Weerasooriya *et al.*, 2023), player disagreement was treated analytically as an indicator of diverse views on alt text suitability or uncertainty, aligning with findings on the variance of expertise in the case of alt text (see Chapter 8). Several measures were as such used to analytically assess such cases of disagreement, including standard deviation, outlier detection, and deviation from the group mean. Including singly rated items could introduce greater noise and reduce the interpretability of the findings. Accordingly, first, players whose ratings had no variation with one another (SD = 0) were examined to filter noisy rating scores. Three such players were identified with each having provided at least 10 rating scores; as such, the first researcher provided a moderation rating score for each alt text description they had rated, and players' rating scores were excluded only if 70%[14] or more of their rating scores were inconsistent with the moderated rating score (>2 deviation). The moderation by the researcher was merely procedural, rather than interpretative, as the researcher only evaluated whether the alt text adhered to instructions from the in-game tutorial (e.g., altC template, functionality mentions for functional images like links or buttons), and was only used in the aforementioned cases where player disagreement was high to ensure tutorial instructions were applied. As a result, all three players' rating scores were excluded from the subsequent analysis.

[14] The choice of more than simple majority threshold (70%) was chosen to avoid exclusion of minor disagreements among player rating scores, accounting for the fact that these players were not experts in alt text annotation. This choice is also in line with peer studies on consistency thresholds in crowdsourcing approaches (Kim *et al.*, 2024).

Then, a further analysis to the alt text descriptions with at least three rating scores took place to identify potential outliers within these ratings (outlier rating scores) for each alt text description compared to the non-outlier ratings for the same description (core group rating scores) (see Fig. 17 below). Alt text descriptions with at least three ratings were selected to enable a more reliable estimation of rating variance and the detection of outliers. A minimum of three data points is the smallest group size that allows for a non-trivial measure of variance and gives an initial sense of consensus or disagreement among raters. This threshold was thus utilized to help reduce the influence of anomalous individual ratings while retaining a meaningful subset of data for quality and agreement analysis.

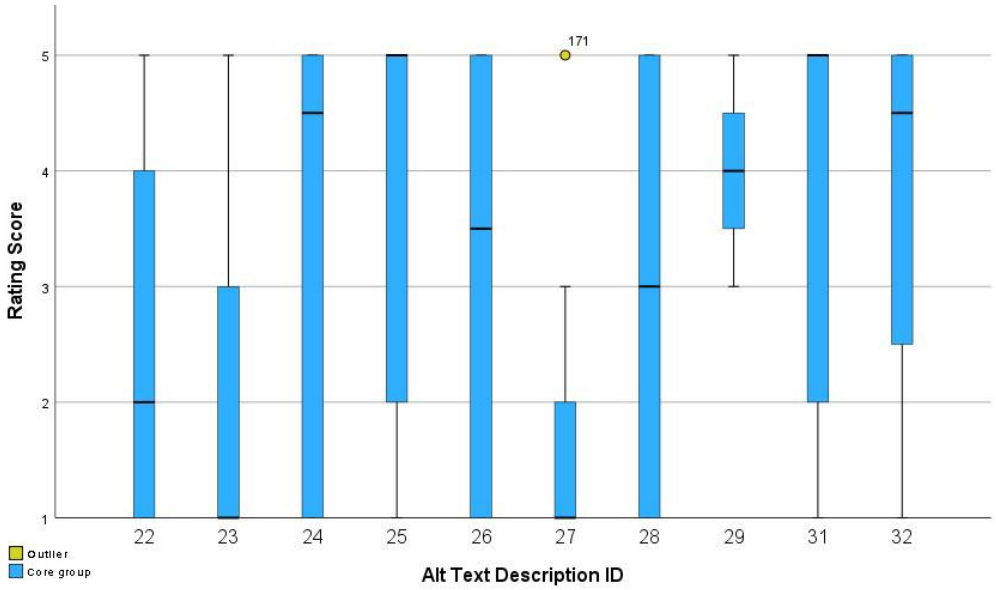


Fig. 17. Example boxplot for identifying outlier rating scores in alt text descriptions with at least 3 ratings

Outlier rating scores (labelled with their non-identifying case numbers (e.g., 171) in the above figure to maintain participant anonymity) were excluded if they were inconsistent with the mean rating score of the core group for each alt text description (>2 deviation from the latter) and were preserved in all other cases. Core group rating scores were also excluded when the outlier rating score was perceived as more accurate than the former and there was again a deviation greater than two between the mean rating score of the core group and the outlier. For example, in the case of the alt text description (ID = 27) in the above figure, core group rating scores range from one to three (Core group mean rating score = 1.33, SD = 0.82), with one outlier rating score of five (case number = 171) excluded due to inconsistency with the former following accuracy evaluation by the first researcher.

Finally, alt text descriptions with exactly two rating scores were also investigated to supplement the analysis. Although limited in depth, this subset allowed for assessing basic

agreement/disagreement and was used to flag descriptions where user ratings diverged most strongly. Initially, the difference between the two scores was calculated, and in cases where the deviation between the two ratings was greater than two, this was considered a substantial disagreement. To resolve these cases and maintain consistency in the analysis, a moderation rating was provided by the researcher, who reviewed the alt text and the associated image. In such cases, the mean rating score between the two players' scores was calculated and the scores were excluded if this mean rating score was inconsistent with the moderation (>1 deviation from the moderation rating score). For clarity, a stricter exclusion threshold was applied (>1 deviation), as mean scores are more susceptible to extreme values, as opposed to individual rating scores (>2 deviation), owing to players not being experts in alt text annotation and limitations of the ordinal rating scale (*e.g.,* scores 1-3 or 3-5 reflecting marginal distinctions). Ultimately, 128 rating scores and 44 alt text descriptions (excluded for no longer having two or more rating scores) were excluded, resulting in a cleaned dataset comprising 386 alt text descriptions and 1410 rating scores. The cleaned dataset was used for the training of the AI models proposed in Section 9.2.3.

However, it must be noted at this stage that a primary limitation in calculating traditional inter-rater reliability (IRR) statistics in this work has been identified in this process due to the low number of raters per item. Specifically, many alt text entries received ratings from fewer than 10 different players, with some having only two or three. Standard IRR coefficients such as Krippendorff's Alpha, Cohen's Kappa, or ICC rely on sufficient ratings per item to produce stable and meaningful estimates. This is corroborated by methodological discussions around sample sizes for IRR coefficients (StackExchange, 2017), and general guidelines on IRR (Gwet, 2014), which advocate that samples sizes need to be predetermined based on parameters, such as desired precision and expected agreement, to avoid compromising the estimated reliability. This is hence a limitation of this study, as the number of raters per alt text description varied and it was as such not possible to meet statistical prerequisites. In such sparse-rating conditions, therefore, these metrics can become unreliable or overly sensitive to small variations. To address this, in this work, alternative descriptive measures of agreement were considered, such as the standard deviation and mean absolute deviation of ratings per item, the proportion of ratings within ±1 star of each other, and a semantic similarity analysis for alt text descriptions with at least two ratings.

10.4.3 Semantic similarity of context prompt impact on quality (S-RQ5)

The correlation between the use of structured context prompts based on the definition proposed in this thesis (Section 4.3.1) and the quality of player-authored alt text was investigated next. To achieve this, the researcher investigated the binary presence of the context definition elements in player-authored alt text descriptions; i.e., each alt text description was investigated in tandem with the context prompt that was presented to players and whenever elements of the context prompt were present/not present in each description, a binary value of one/zero was assigned to that element, respectively. All assigned binary values were summed up into a total named the presence score of each alt text description, ranging from zero (no presence of context prompt elements) to five (all five context prompt elements were present). Importantly, it is clarified that the researcher did not rate the descriptions; these were only checked for the binary presence of context elements based on the context definition (Section 4.3.1). Accordingly, the presence scores calculated were used in three levels of analysis to examine the:

1. Overall relationship between the context prompts and the quality of player-authored alt text;
2. Relationship between increasing context presence scores and quality;
3. Relationship between image-specific context elements (type, function, intent) and quality, as well as webpage-specific context elements (topic, purpose) and quality.

Fig. 18 below refers to the first level of analysis, showing the overall relationship between the presence score and the mean rating scores given by players.

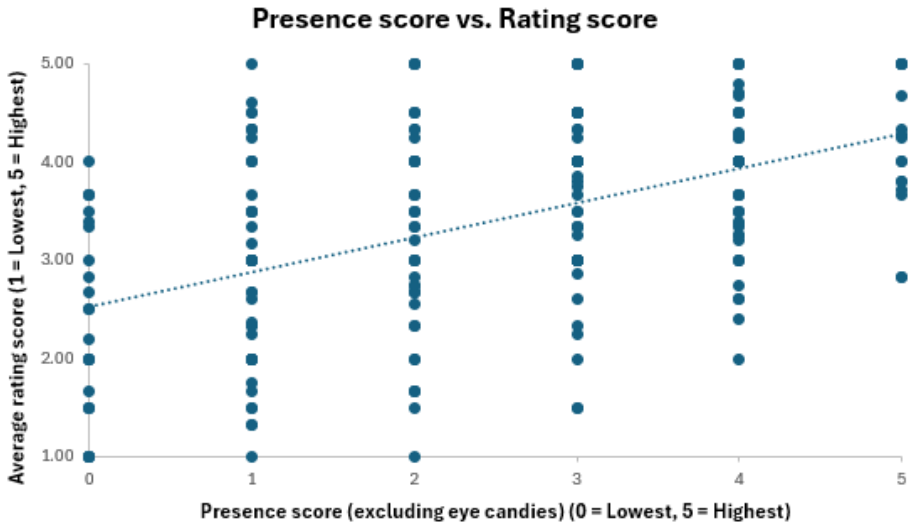


Fig. 18. Correlation between the binary presence score of context prompt elements in player-authored alt text descriptions and average rating scores

The above figure shows a positive monotonic relationship between higher presence scores relating to higher quality perceptions of alt text by players, who have generally given higher rating scores to alt text descriptions with higher presence scores. Spearman's rho was additionally calculated (0.467, $p < .001$) further supporting a positive, moderate, and statistically significant correlation between context presence score and mean rating score.

Then, the validity of this correlation was explored by incrementally investigating the binary presence of context elements in alt text descriptions to determine whether an increase in presence score was related to increased mean rating score and decreased standard deviation (see Table 11 below).

Table 11. Impact of incremental presence of context elements in alt text descriptions on their perceived quality

	Presence score = 0	**Presence score = 1**	**Presence score = 2**	**Presence score = 3**	**Presence score = 4**	**Presence score = 5**	**Average**
Number of alt text descriptions	26	45	44	55	46	17	N/A
Mean rating score	2.31	2.94	3.29	3.68	3.84	4.19	3.37
Mean increase		0.63	0.35	0.39	0.16	0.35	0.38
Standard dev.	1.38	1.31	1.19	1.10	1.11	0.75	1.14
Std. dev. decrease		0.07	0.12	0.09	-0.01	0.36	0.13

The table is revealing in several ways. First, it further lays claim to the notion that higher presence score relates with higher quality perceptions of alt text by players, revealing a consistent positive trend, with the mean rating score increasing by 0.38 on average with each increment in presence score. The table also indicates a higher level of consensus between player rating scores with each increment in presence score, with standard deviation decreasing by 0.13 on average with each such increment. These trends, therefore, further support the notion that

the provision of structured context prompts can help players author alt text descriptions perceived as of higher quality.

Finally, a third level of analysis to further investigate the impact of the structured context prompts on quality perceptions of alt text by players was conducted, involving descriptions with binary presence of all image-specific (type, function, intent) and all webpage-specific (topic, purpose) elements of the proposed context definition (see section 9.2.1.3). Table 12 below compares the mean rating score, standard deviation, and the correlation coefficient (Spearman's rho) and its statistical significance between all alt text, image-specific, and webpage-specific alt text descriptions

Table 12. Impact comparison of overall, image- and webpage-specific context elements on alt text description quality

	Overall	**Image-specific**	**Webpage-specific**
Count	233	46	56
Mean rating score	3.38	3.74	4.09
Standard deviation	1.16	1.15	0.86
Spearman's rho	0.467	0.475	0.135
Statistical significance	$p < .001$	$p < .001$	$p = .352$

The table reveals that alt text descriptions that incorporated all webpage-specific elements of the context definition had the highest mean rating score (4.09) and the highest consensus among raters (SD = 0.86). Although this indicates higher quality perception of descriptions incorporating webpage-specific elements, the correlation coefficient for these descriptions was weak (Spearman's rho = 0.135) and not statistically significant ($p = .352$). Thus, and despite what players' rating scores suggest, the incorporation of webpage-specific elements in alt text descriptions is not a reliable indicator of the higher quality of such descriptions. Contrastingly, alt text descriptions incorporating all image-specific elements also had a higher mean rating score (3.74, SD = 1.15) than the overall mean rating score (3.38, SD = 1.16), albeit more reliable than descriptions incorporating all webpage-specific elements, owing to the high correlation coefficient (Spearman's rho = 0.475) and statistical significance ($p < .001$). Similar correlation (Spearman's rho = 0.467) and statistical significance ($p < .001$) is observed in alt text descriptions overall, indicating the significant impact of context prompts elements in quality perceptions of alt text by players, not least in relation to descriptions that incorporate all image-specific elements of the context definition.

Taken together, these three levels of analysis of the impact of structured context prompts on quality perceptions of alt text by players revealed that the incorporation of such context prompts improves the perceived alt text quality. This improvement was moderate and

statistically significant (0.467, $p < .001$) for alt text descriptions with at least two rating scores, and an increase in the mean rating score and consensus between players was observed with each increment in presence score. Alt text descriptions incorporating all webpage-specific context elements (topic, purpose) achieved the highest mean rating score and agreement, but the correlation was weak and not statistically significant. The strength of the correlation was, in fact, most evident in descriptions incorporating all image-specific elements. Thus, it can be surmised that the higher the binary presence of context prompt elements, the higher the impact of structured context prompts on quality perceptions of alt text descriptions, not least for descriptions incorporating all image-specific elements of the proposed context definition.

10.4.4 Descriptive verbosity and quality (S-RQ6)

To address this research question, the textual length (in character count) for each authored alt text description was used as a proxy for verbosity. While this does not directly capture any richness in terms of semantic or informational content, it serves as a useful first-order approximation of how detailed a description might be and how this may correlate with perceived quality. Accordingly, correlation analysis was used to investigate correlations between alt text length and mean rating score to determine if more verbose alt text descriptions (authored by players) relate to a higher quality perception of the alt text. Initially, a point-biserial correlation was used to see whether there was any correlation between decorative images (eye candy) and players' ratings in terms of whether images having been marked as eye candy impact the quality perception of alt text (see Fig. 19 below).

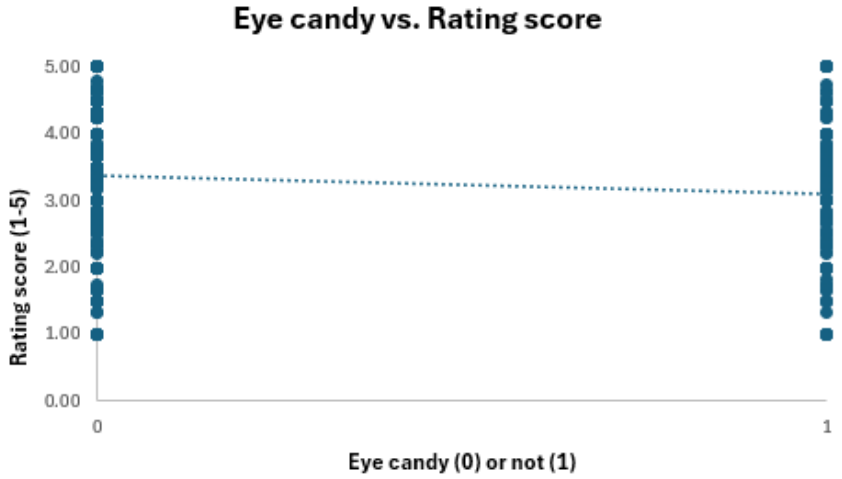


Fig. 19. Point-biserial correlation between decorative images (eye candy) and player rating scores

The analysis revealed a point-biserial correlation of -0.13, which suggests that there is no meaningful relationship between an image having been marked as eye candy and players' rating scores. As such, images that were marked as 'eye candy' by players were excluded from this analysis. Following that, the correlation between alt text length (character count) and players' ratings using Spearman's rho was calculated (0.54), owing to the ordinal dataset and the presence of outliers, revealing a positive monotonic relationship between the two variables (see Fig. 20 below), which shows that more verbose descriptions relate to higher quality perceptions of alt text.

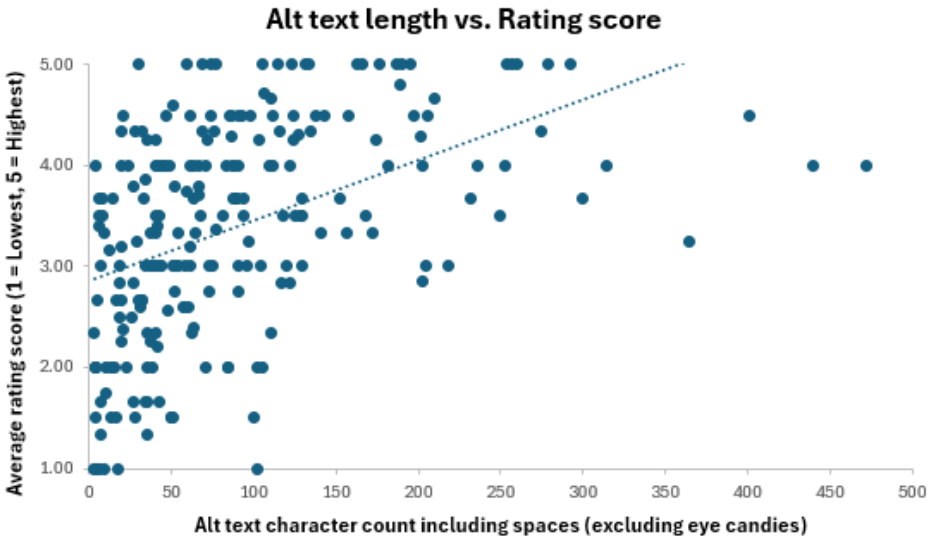


Fig. 20. Correlation between the character count of alt text descriptions and player rating scores

As can be observed in the above figure, whilst a level of variability exists across different rating scores, the trendline reveals an overall positive correlation. This indicates that, despite

individual variations in rating scores, alt text descriptions with higher character count are perceived as of better quality and are generally rated higher by players. To further investigate the validity of this correlation between verbosity and quality perceptions of alt text, five high rated (mean rating score between 4 and 5) and low rated (mean rating score between 1 and 2) alt text descriptions were examined, as shown in Table 13 below.

Table 13. Comparison of high- and low-rated alt text descriptions on the correlation between verbosity and quality

Category	Sample description	Character count	Num. of rating scores	Mean rating score	Standard dev.
High-rated alt text descriptions	A Starbucks cup printed with friendly Christmas themed line art, on a ledge with a blurred cityscape in the background. There's an inviting drip of latte on the rim.	166	5	5	0
	A man in a wetsuit relaxes on the beach after a surfing session, his double-finned surfboard leaned against his head to block the sun, watching the waves roll in past the crags	176	4	5	0
	The archipelago of small islands seen from birds perspective, different sizes, islands have a lot of vegetation, the central island has a castle	144	6	4.5	0.5
	Bar chart showing that annual circulation of Fantastic was notably highest during 1964-1967 | Detailed Description Whereabouts	127	10	4.3	0.78
	Surfer with the right gear and board, staring at the waves he'll tame | Link Destination	88	4	4	0
Low-rated alt text descriptions	|Button operation	17	5	1	0
	Settings	9	5	1	0
	Cofffee	7	6	1.33	0.75
	Simple outline illustration of a flying aeroplane .	51	4	1.5	0.5
	A wonderful photo from a place that someone can go	50	4	1.5	0.5

The character counts of alt text descriptions are evidently higher in the high-rated descriptions (Avg. = 140.2) compared to the low-rated descriptions (Avg. = 26.8). These values are also quite reliable considering that the former were rated by 5.8 different raters on average, while the latter were rated by 4.8 raters on average, not to mention that the average deviation for their rating scores was 0.26 and 0.35 for highly and lowly rated descriptions, respectively. These

numbers, thus, corroborate the indicated throughout this section strong correlation between the verbosity of alt text descriptions and players' quality perceptions.

10.4.5 Player rating consistency (S-RQ7)

Finally, it was essential to also determine if there are any agreements or disagreements between players in terms of rating consistency. Owing to the limitation discussed in Section 10.4.2, an alternative descriptive measure of agreement was employed. First, the rating score distribution across the player population for descriptions with at least two ratings was examined. As shown in Fig. 21 below, 908 (64.4%) alt text received a positive (>= 3) rating score, which, if taken together with the mean rating score (average rating score across alt text with at least two ratings) of 3.27 (SD = 1.24), indicate a general positive perception of the suitability of alt text authored by players.

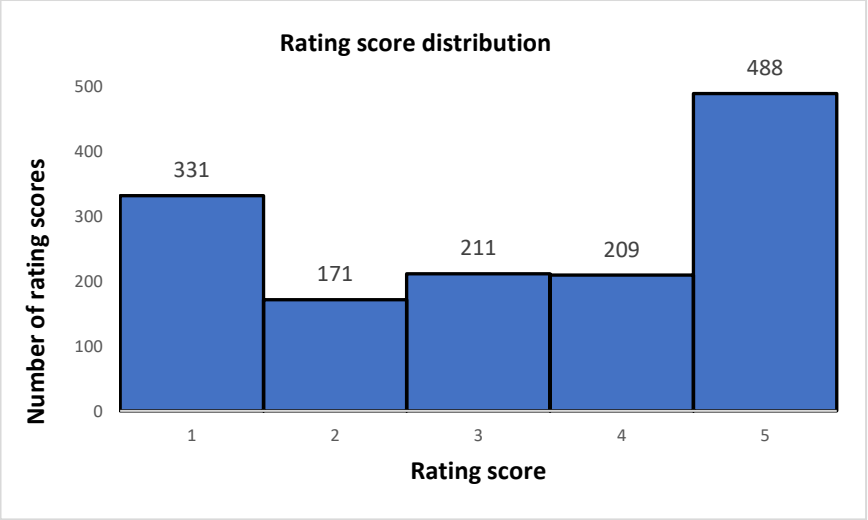


Fig. 21. Distribution of rating scores for player-authored alt text descriptions with at least 2 ratings

Then, in order to determine the level of consistency among players, a mean absolute deviation (MAD) of 0.95 was calculated, as it is less sensitive to outliers than standard deviation. Next, the proportion of raters whose rating fell within ±1 of the mean rating score for each alt text description was examined (see Fig. 22 for a distribution of mean rating scores across all alt text descriptions), which revealed a good level of agreement (Avg. = 65.28%) across player ratings.

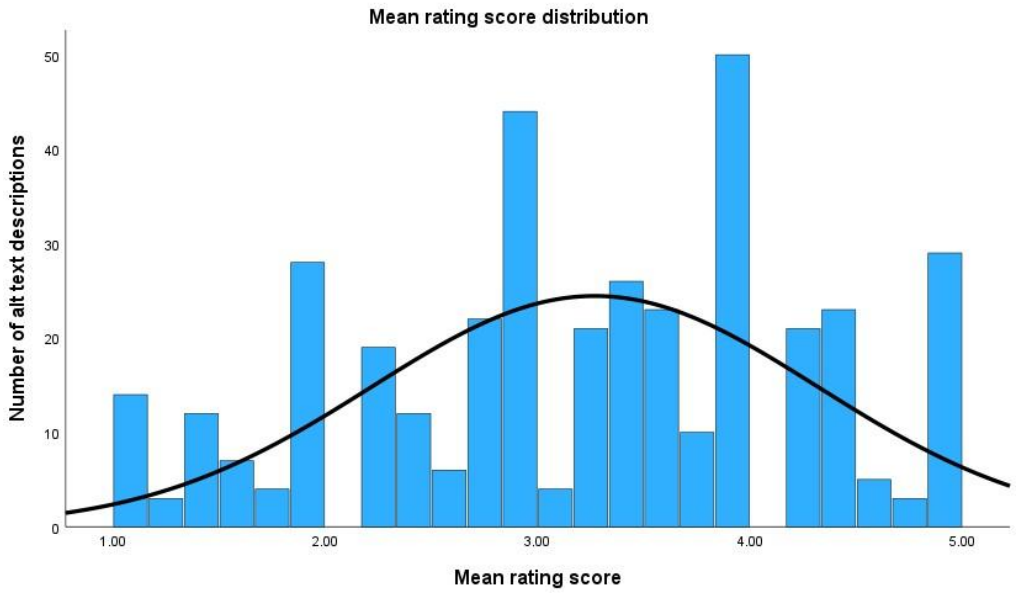


Fig. 22. Distribution of mean rating scores across player-authored alt text descriptions with at least 2 ratings

The previously mentioned level of agreement (Avg. = 65.28%) was observed across alt text descriptions with at least 2 rating scores (n = 386) and mean rating scores ranging from one (SD = 0) to five (SD = 0), with an average mean rating score of 3.27 (SD = 1.02), as depicted in Fig. 22. Taken together, these findings suggest that there is a positive level of consistency among player rating scores, further suggesting the usefulness of our approach.

10.5 Chapter summary

Following the presentation of TagALTlong, a novel context- and community-driven GWAP approach to alt text annotation and evaluation, this chapter investigated its effectiveness (Section 10.4.1), as well as its perceived quality (Sections 10.3.3 and 10.3.4) and consistency (Section 10.4.5) via a mixed-method analysis of results based on crowdsourced annotations, user ratings and consensus measures. Overall, the findings suggest that TagALTlong is an effective approach to generate suitable alt text annotations at scale, grounded in the finding that verbose descriptions are perceived as higher quality (Fig. 20 and Table 13). The findings also revealed that the incorporation of a context-based prompt in the annotation process related with higher quality perceptions of alt text by evaluators (Fig. 18 and Tables 11 and 12). These results are significant when seen against the discussed limitations in the suitability of automatically generated alt text through SoTA V2L captioning models (Chapter 6), where training data derive largely from web scraping and thus lack context. Whilst it was shown that very large datasets are needed to train such models, the findings of this study highlighted the effectiveness of the GWAP approach to collect context-driven alt text data at scale, while also underscoring the

impact of context on the quality of alt text. It is therefore now important to also investigate the potential of the GWAP-generated dataset as an alternative to web-scraped data for training V2L models and evaluate the performance of automation through ML models in terms of accessibility (i.e., alt text suitability and context presence in automatically generated alt text), rather than captioning benchmarks.

Whereas automating alt text generation through AI is essential to address the reluctance to author alt text and the increased abundance of multimedia content on the Web, a key gap is the lack of evaluation of the performance of SoTA V2L models on accessibility (Section 4.4.1 and Section 6.3). Accordingly, Chapter 11 presents the evaluation of the performance of the AI models developed in this thesis (Chapter 9) in terms of alt text quality and the ability to generate context-driven alt text when fine-tuned and trained on the GWAP-generated dataset compared to pure image processing.

Chapter 11. Model performance evaluation

11.1 Introduction

The previous chapters presented the design of TagALTlong, the first GWAP for context-driven alt text annotation and evaluation (Chapter 9), and the results of a user study where players were recruited to play TagALTlong to gather a dataset of alt text descriptions and rating scores (Chapter 10). Results highlighted the effectiveness of the GWAP approach to gather alt text data at scale, as well as the important role of context in the quality of alt text. The resulting dataset of player-authored alt text descriptions and rating scores that incorporate context in alt text is an important contribution given the lack of similar datasets and crowdsourcing-based approaches and the lack of context in current efforts. Although the game can stand on its own as a solution to alt text barriers, it was discussed that automating alt text generation is essential to address the reluctance to author alt text and the need to scale alt text generation on par with the increase in multimedia content on the Web (Chapter 6). Two AI models were therefore developed and used in this thesis to automatically generate alt text descriptions (Chapter 9), and this chapter presents the evaluation of the performance of these models in terms of the quality of the automatically generated alt text and the presence of context in alt text, addressing a reported gap in the evaluation of the performance of models on accessibility (Section 4.4.1 and Section 6.3). The twofold evaluation of the models in this chapter addresses the following sub-research questions (S-RQs):

- S-RQ8: Is the use of a GWAP-generated dataset to train the AI model (trained model) for generating alt text descriptions an effective approach to get closer to a human average compared to pure image processing (control model)?
- S-RQ9: How well does the trained AI model classify decorative (eye candy) images?
- S-RQ10: To what extent does learning from structured context prompts improve context-driven alt text generation?

Based on the S-RQs, an empirical evaluation is conducted; first, via a user study to assess whether it is possible to approximate average human-level alt text quality while automating the alt text generation process (S-RQ8, S-RQ9); and, second, via an investigation of the binary presence of elements of the context definition (see section 4.3.1) in automatically generated alt text descriptions (S-RQ10).

11.2 User study evaluation (S-RQ8, S-RQ9)

To address S-RQ8 and S-RQ9, a comparative user study was conducted between the AI model that was Google's BERT (for classification) and T5-small (for caption generation), which were fine-tuned and trained on the GWAP-generated data (human-curated alt text descriptions and rating scores) and the T5-small model (see Fig. 13). Both models generate captions from the same input which is formed of image processing data and image context. The same generation prompt was used for both models. The primary aim of the study was to assess whether the AI model can approximate average human-level alt text quality (S-RQ8). The study further aimed to investigate how well the model trained on the GWAP-generated output distinguishes between decorative and non-decorative images (S-RQ9), making it the first model to attempt this crucial pathway towards improving web navigation via screen readers. The study design, data collection, and analysis procedures are detailed in the following subsections.

11.2.1 Study design, sampling strategy and procedure

Ethics approval was granted by the researcher's institutional Research Ethics Committee (Ref: 41665-A-Feb/2025- 53701-1) (see Appendix G). The study adopted a quantitative, exploratory design to evaluate the impact of the context-driven GWAP output (alt text descriptions and peer ratings) on participants' quality perceptions of alt text. In this vein, participants were asked to complete an online survey where they rated (using a 1-5 Likert scale) a sample of alt text descriptions generated by the trained and the control models for the same image-context pairs (see Appendix F for a sample of survey format). The sample comprised alt text descriptions for 20 image-context pairs generated by both the control and the trained models, resulting in 40 image-context-alt text tuples. The sampling criteria were as follows:

- Prioritise alt text descriptions with the most rating scores in TagALTlong.
- No duplicate images or contexts.
- A good mix of context prompt elements.
- 30% of the image-context pairs to have been marked as 'eye candy' (decorative images) in the GWAP to investigate the model's ability to classify decorative images.
- A close population distribution

For trustworthiness, the alt text descriptions that had received the most rating scores by players of the GWAP were selected. To avoid redundancy, no image or context prompt was used more than once, and the selected context prompts presented a variety of context prompt elements

(e.g., links, logos, non-functional images, or images found on a health/social media webpage). In line with S-RQ9, 30% of the image-context pairs selected for the sample had been marked as eye candy by players of the GWAP to show the potential of the model in distinguishing between decorative (eye candy) and non-decorative images. The objective was also to use a sample that represented the population well; therefore, the distribution in the sample aimed for a difference within two decimal points with the distribution in the population, while average mean rating score was deemed more important than standard deviation for this difference (see Table 14 below). As can be observed in this table, the sample is a close representative of the population, as a difference within one decimal point was achieved.

Table 14. Distribution difference between sample and population

	Overall mean rating	**Overall std. dev.**	**Non-decorative mean**	**Non-decorative std. dev.**	**Decorative mean**	**Decorative std. dev.**
Sample	3.2	0.63	3,16	0.66	3.29	0.58
Population	3.23	0.72	3.24	0.81	3.19	0.58

The survey adopted a within-subjects design, where every participant rated all 40 image-context-alt text tuples (20 unique image-context pairs; each pair had one alt text description generated by the control model and one by the trained model). The survey was administered via Microsoft Forms; specifically, participants were provided with a private link to access the survey on this platform. In line with the objective of the study, participants were able to access and complete the survey in a naturalistic, unsupervised setting in their own time and using their own devices. No manipulation or experimental intervention was introduced; instead, evaluation relied exclusively on data captured via survey completion sessions. All data were generated exclusively via participants' survey completions. Upon visiting the survey, they were first prompted to read the Participant Information Sheet (PIS) and Consent form (see Appendices H and I, respectively), and following that, they were asked to provide electronic Consent by agreeing to take part in the study. Only participants who provided consent were able to continue and interact with the survey. At the top of all pages of the survey, participants were provided clear instructions on how to rate each alt text description for suitability based on each image and its specific context, which were the same instructions found in TagALTlong. On average, each survey session lasted approximately less than 20 minutes, depending on the participant.

11.2.2 Participants

The recruited participants were a subset of participants from the original pool of players who evaluated TaglALTlong (see section 10.3). Although these individuals are not domain experts,

they had previously authored and evaluated alt text descriptions as part of the initial data collection phase. Their prior involvement ensured familiarity with the goals of the task and with typical qualities of suitable alt text. Further, using a small, purposive sample made it possible to obtain focused, comparative judgements between human-authored and model-generated outputs, leveraging participants' existing knowledge of the task at hand. This is in line with established practice in human-centred AI research, where depth of insight and task-specific familiarity are often prioritised over large sample sizes when the goal is formative evaluation (Shneiderman, 2022) and it aligns with past research supporting the use of small, representative samples when the objective is to assess whether outputs meet users' expectations (Nielsen, 2000). Further past work, however, heeds the unreliability of usability testing with user samples comprising five or less participants, demonstrating clear improvement rates with larger samples (Faulkner, 2003). Therefore, in this work, the focus will be to transparently outline the criteria relating to the selection of the sample, whilst aiming for a representation that closely aligns with demographic characteristics of the population. It is nonetheless noted that findings are as such not a definitive interpretation of the broader population. Participation was voluntary and no financial incentives were offered. The researcher contacted former TagALTlong players, who had provided consent to be contacted for follow-up studies upon registering on TagALTlong (see section 10.2). These players were invited via email to participate in this online survey, while aiming for a sample that closely represented the population in terms of gender and age distribution (see Table 15 below). Seventeen (N = 17) unique former TagALTlong players (13.6%) accepted the request to participate in the survey in June 2025.

Table 15. Gender representation and age average difference between sample and population

	Age average	**Male representation**	**Female representation**	**Prefer not to say representation**	**Other representation**
Sample	29.76	10 (58.82%)	6 (35.29%)	1 (5.88%)	0 (0)
Population	29.79	68 (60.71%)	39 (34.82%)	4 (3.57%)	1 (0.89%)

The above table shows a fair per-gender representation of the population in the selected sample of participants and a difference within one decimal point in terms of average participant age. For clarity, gender has been used in accordance with the definition of the WHO (2025).

11.2.3 Data normality and distribution

First, the distribution of the data was evaluated based on the assumption of normality to determine the appropriate statistical tests (parametric/non-parametric) for ensuring the validity

of the findings. The study participants (N = 17) rated a total of N = 680 alt text descriptions split across two independent groups (Control: N = 340 rating scores; Trained: N = 340 rating scores) for 20 unique image-context pairs. To evaluate normality, the Shapiro-Wilk (Control – W: .877, $p < .001$; Trained – W: .904, $p < .001$) and Kolmogorov-Smirnov (Control – W: .177, $p < .001$; Trained – W: .174, $p < .001$) tests were used; the tests confirmed significant deviations from normality in both groups, and were further supported by visual inspections (Fig. 23).

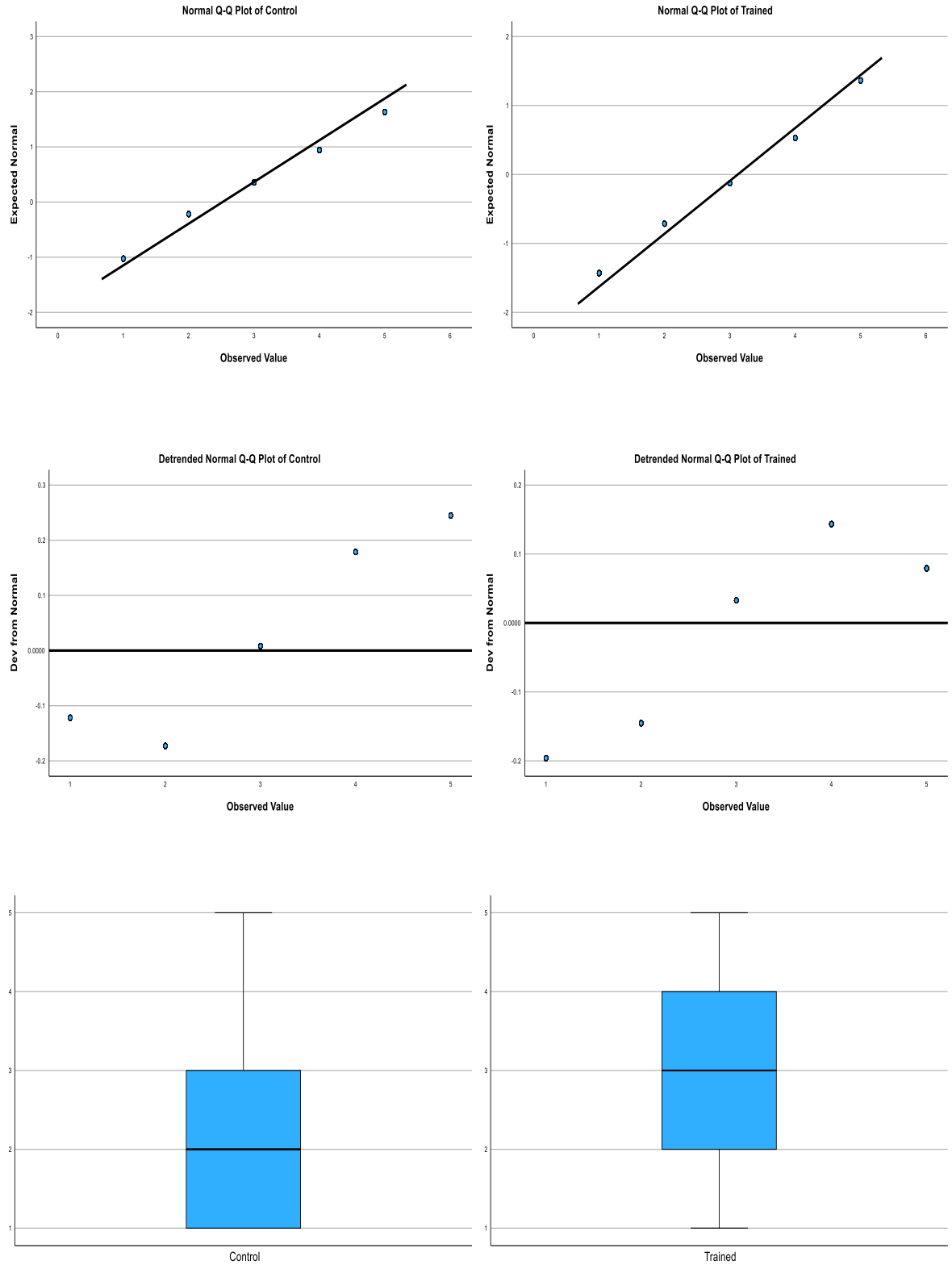


Fig. 23. Normality and distribution plots. (a, b) Normal Q-Q plots control (top left: control; top right: trained); (c, d) Detrended Q-Q plots (middle left: control; middle right: trained); (e, f) Boxplots (bottom left: control; bottom right: trained)

The data were also not normally distributed at the image-context pair level, with 16 and 14 out of the 20 image-context pairs showing significant deviations from normality ($p < .05$) in the

control and trained groups, respectively. Therefore, non-parametric statistical tests were in order for both the overall data and the paired testing at the image-context pair level, where the alt text generated by the trained model for each unique image-context pair was compared to its control model counterpart.

11.2.4 Training effectiveness on generated alt text quality (S-RQ8)

The performance of the model that was fine-tuned and trained on the GWAP-generated dataset compared to pure image processing in terms of approximating human-level quality alt text was investigated next. Specifically, the null hypothesis tested was as follows: 'The distribution of rating scores is the same across categories of Control and Trained'. To achieve this, the Mann-Whitney U test was used to compare the aggregated rating scores across the 20 image-context pairs in both groups (control and trained) independently (see Fig. 24 below).

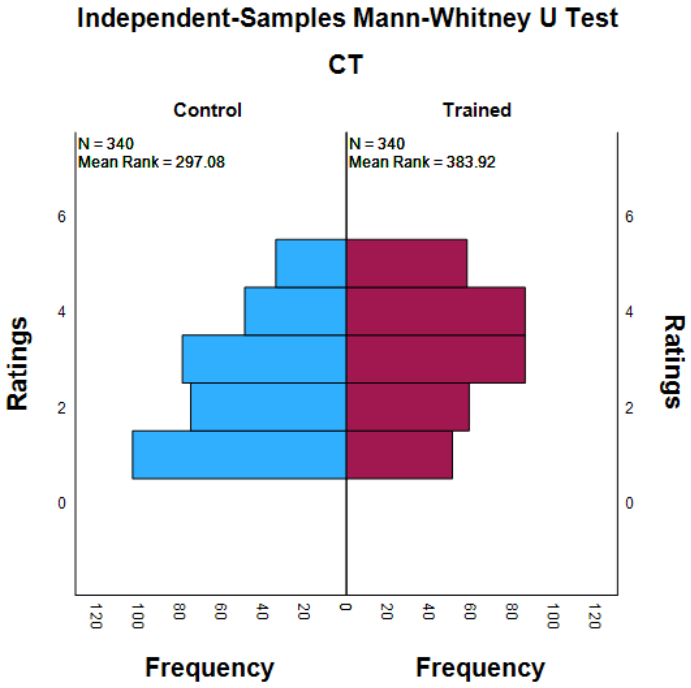


Fig. 24. Distribution of rating scores across the control (left) and the trained (right) groups

The above figure reveals significantly higher rating scores in the trained group (Mean rank = 383.92) compared to the control group (Mean rank = 297.08), indicating the strong, positive effect of training on the perceived quality of generated alt text descriptions. The null hypothesis was, thus, rejected by the independent-samples Mann-Whitney U test ($U = 72{,}564$, $p < .001$), and a further null hypothesis was tested; i.e., 'The median of differences between Control and Trained equals 0', using a related-samples Wilcoxon Signed-Rank test (see Fig. 25 below).

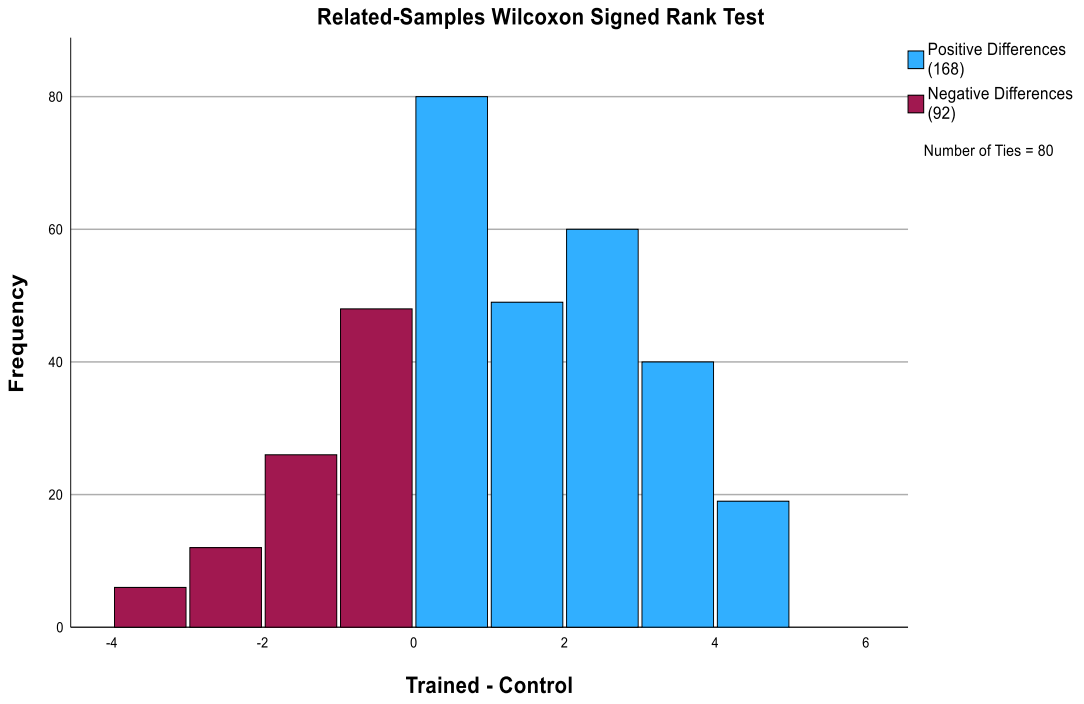


Fig. 25. Distribution of rating score differences between the control and the trained groups

The distribution of differences shown in the above figure rejects the null hypothesis; i.e., the median of such differences equals 0; instead, revealing a strong, positive skewness, with 168 positive differences (higher rating scores in Trained) compared to 92 negative differences (higher rating scores in Control). Eighty ties, i.e., identical rating scores for the same image-context pair for both the alt text generated by the trained and the control (baseline) model, were observed, highlighting identical low rating scores across the two groups. These results indicate a significant net improvement rate of 64.6% (ignoring ties) between alt text descriptions generated by the trained model compared to the control model, highlighting the improved performance of the former, which was fine-tuned and trained on the GWAP-generated dataset, subsequently demonstrating the value of the dataset. Therefore, the related-samples Wilcoxon Signed-Rank test supports the results of the independent-samples Mann-Whitney U test, indicating significantly higher rating scores in the trained group ($Z = 5.803$, $p < .001$). However, it must be noted that although these tests were non-parametric, aligning well to the data not being normally distributed (see previous section), a limitation with non-parametric tests is that they make no assumptions about the distribution of the population. This was addressed in this evaluation by the selection of a sample that is representative of the population based on the strategy outlined in Section 11.2.2.

Accordingly, the effectiveness of training was also investigated at the image-context pair level, i.e., the rating scores for the alt text descriptions generated by the control model for each

image-context pair (C01-C20) were compared to their counterparts from the trained model (T01-T20). To achieve this, Cohen's d effect sizes were calculated for each CT pair to measure the magnitude and the direction of differences between quality perceptions of the two models' outputs. These training effect sizes were mapped into a forest plot alongside their 95% confident intervals (CI) for each pair, which are included for replicability (see Fig. 26).

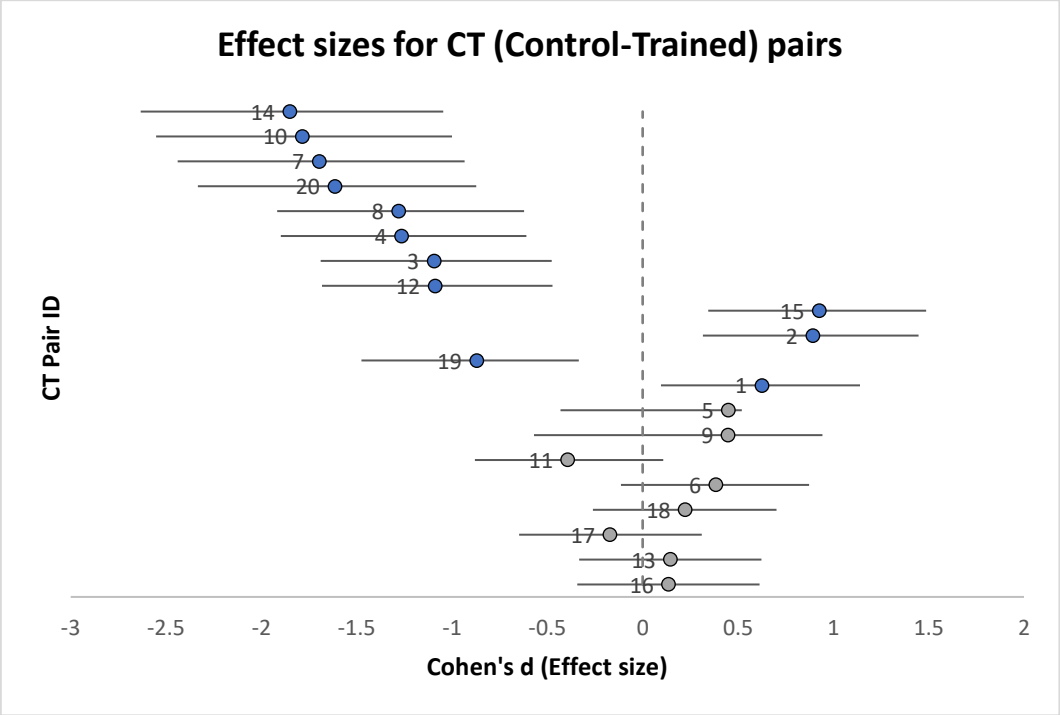


Fig. 26. Forest plot of Cohen's d effect sizes for CT (Control-Trained) pairs with error bars representing 95% CIs (sorted by absolute magnitude). Color-coded for statistical significance (blue: $p < .05$; grey: $p \geq 0.5$)

The above plot is revealing in several ways. First, CT pairs are sorted from strongest to weakest absolute Cohen's d effect size values (absolute magnitude); thus, the bigger the difference between the control and trained model, the higher the CT pair is shown in the plot. Second, the line of no effect (i.e., the vertical barred line at zero in the plot) allows for a clear interpretation of when the perceived quality of the output of each model outperformed the other. Specifically, the trained model outperforms the control model when Cohen's d is negative (below zero) (11 cases), and vice versa (above zero) (9 cases). Further, the Wilcoxon Signed-Rank test, which is well-suited for paired data, was used, indicating statistically significant ($p < .05$) and non-significant ($p \geq 0.5$) differences in rating scores for alt text descriptions generated by the trained and control models. Twelve out of twenty cases were statistically significant (blue circle points in Fig. 26 above), with nine of those being cases where the trained model outperformed the control model. Thus, the trained model substantially improved alt text descriptions generated by the control model in 75% of statistically significant cases (as per participants' quality

perceptions), highlighting the effectiveness of using a human-curated output to train such models for generating better alt text.

More colloquially, in the context of accessibility, effect sizes ($p < 0.5$) being preferred in 75% of statistically significant cases (9 out of 12 cases) refers to alt text descriptions that can be reliably accessed via screen readers by VIUs. Effect sizes ($p \geq 0.5$), on the other hand, show lack of context and human-curation, which appears to reflect on the perceived quality of alt text by human raters, as such descriptions were only preferred in three statistically significant cases. Therefore, the use of human-curated, context-driven data for V2L model training is most beneficial for accessibility when the control output equivalent is of low quality (rating score 1-2), with the relevant effect sizes revealing that the dataset helps transform otherwise alt text descriptions that would have otherwise been irrelevant for accessibility into description that VIUs can very relevantly (rating score 4-5) access via screen readers. To further investigate the validity of the indicated magnitude and direction of the difference between quality perceptions across CT pairs, the three strongest Cohen's d effect sizes where the trained model was preferred over the control model and vice versa, as well as three illustrative non-statistically significant cases, were examined (see Table 16 below).

Table 16. Comparison of CT pairs with the most notable model performance differences

Category	Pair ID	Control alt text	Trained alt text	Control mean	Trained mean	Wilcoxon Z	Cohen's d	Significance (p)
Trained > Control	CT14	person, tree, sky	A picture of a young woman holding a watermelon slice in front of her face.	1.59	3.94	-3.443	-1.850	< .001
	CT10	two bottles of vitamin c3	vitamin c and b12 tablets	1.82	3.71	-3.555	-1.786	< .001
	CT07	person, bicycle, umbrella, road, sky	a man riding a bike in a valley	1.59	3.41	-3.360	-1.697	< .001
Control > Trained	CT15	woman standing on a ledge with her head tilted	a young woman with a head tilted on a ledge infront of a social media page	3.12	2.06	-2.859	.926	.004
	CT02	graph showing the number of students in each class	This photo is marked as 'Eye Candy'.	3.00	1.59	-2.699	.892	.007

Category	Pair ID	Control alt text	Trained alt text	Control mean	Trained mean	Wilcoxon Z	Cohen's d	Significance (p)
	CT01	coffee beans and a coffee pot	A photo of coffee beans showing how to make your own coffee.	3.82	2.94	-2.080	.626	.038
No difference	CT05	zeimm & rosen hand cream	This is a hand cream	2.71	2.65	.000	.045	1.000
Non-significant	CT11	dog, sea, sky, rock	This is a picture of a dog running on the beach with his mouth open.	2.65	3.23	-1.667	-.392	.95
	CT16	black and white photo of a tree	A black and white photo of a tree with a bright blue sky.	3.00	2.82	-.5667	.137	.571

The qualitative data (alt text descriptions) in the above table highlight the most significant improvements in cases where the control model generated descriptions more closely resembling keyword tags than alt text (CT07, CT14) or non-specific descriptions (CT10), wherein the trained model integrated rich contextual information. Conversely, alt text generated by the control model was preferred in cases where the trained model added confusing or irrelevant information that did not match what the image depicted (CT15, CT01) or when the model misinterpreted an important image, such as a graph, for a decorative (eye candy) image (CT02). Whereas it appears that the trained model was preferred over the control model (and vice versa) based on the former's ability to correctly interpret and integrate rich contextual information in the alt text descriptions, it is important to note that the three cases where the control model was preferred were also the only statistically significant cases, as opposed to 9 out of 11 total cases for the trained model being statistically significant. In non-statistically significant cases, it appears that contextual information integrated by the trained model was trivial (CT16) or violated instructions given to participants, such as including 'This is a picture' (CT11). There was also an image-context pair (CT05) where the Wilcoxon test confirmed no statistical difference ($Z = .000$; $p = 1.000$), despite the pair appearing as a case where the control model was preferred in the forest plot (see Fig. 26) based on its Cohen's d effect size (.045). It is also important to note that in cases where the trained model was preferred, the alt text descriptions generated by the control model were lowly rated (1-2 mean rating score). Conversely, in cases where the control model was preferred, the alt text already had acceptable ratings (3-4 mean rating score) before training. Therefore, the effectiveness of training is most

evident when the perceived quality of alt text generated by the control model is poor and it is linked to the trained model's ability to correctly interpret contextual information.

In its current state, the trained model's potential to improve poor quality descriptions generated by the control model has been highlighted and the former was also preferred in 9 out of 12 (75%) statistically significant cases. It is therefore necessary to gather human-curated data (alt text descriptions and rating scores) at a larger scale to train such models to further investigate their potential to improve the quality of generated descriptions.

11.2.5 Classification of decorative (eye candy) images (S-RQ9)

Finally, the trained model's ability to classify decorative (eye candy) images was investigated, making it the first model to attempt to make this crucial distinction between decorative and non-decorative images. To achieve this, the image-context pairs that had been marked as decorative by players in TagALTlong were used to compare with the equivalents generated by the trained model (see Table 17 below).

Table 17. Alt text descriptions generated by the trained model for image-context pairs marked as decorative in TagALTlong incl. context prompts

Pair ID	Trained alt text	Context prompt	TagALTlong mean	Trained mean
T15	a young woman with a head tilted on a ledge infront of a social media page	This photograph is found on a social media webpage with the goal of selling. The intention of the image is to guide (e.g., instructions or steps).	3.46	2.06
T16	A black and white photo of a tree with a bright blue sky.	This painting is a button on a social media webpage with the goal of advertising. The intention of the image is to illustrate (e.g., showing a product).	3.5	2.82
T17	This painting is marked as 'Eye Candy'.	This painting is found on an education webpage with the goal of entertaining. The intention of the image is to complement (e.g., enhancing the text context).	3.56	3.59
T18	This photo is marked as 'Eye Candy'.	This photograph is found on a travel webpage with the goal of advertising. The intention of the image is to illustrate (e.g., showing a product).	3.86	3.3
T19	The image is of a signpost on a street directing to a library. The signpost is on the left and there's a tree on the left.	This photograph is found on an education webpage with the goal of entertaining. The intention of the image is to illustrate (e.g., showing a product).	3.14	2.94
T20	A woman with blonde hair standing in front of an auditorium. Button Operation	This photograph is a button on a health webpage with the goal of informing. The intention of the image is to elicit emotion (e.g., compassion).	2.2	3.4

The table highlights the ability of the trained model to match human classification of decorative images, achieving mean rating scores comparable to human consensus in cases T17 and T18. Further, the model manages to outperform human-authored alt text in case T20, where its classification of a functional (button) image was rated higher than its TagALTlong counterpart, where it had been marked as decorative. These results are deemed reliable, as the sample of image-context pairs that were marked as decorative in TagALTlong were selected for having been rated by the highest numbers of unique raters, and they were subsequently rated by all 17 survey participants. These demonstrate the trained model's ability to classify not only the image itself but also its role in the context it is used in; thus, it classifies decorative images on par with humans (T17, T18), and it improves human classification (T20) by factoring in human consensus through aggregated mean rating scores. However, alt text descriptions generated by the trained model for the remainder of the decorative cases (T15, T16, T19) received lower mean rating scores than their TagALTlong counterparts. These cases highlight a systematic limitation in HumanALT-O-matic's ability to integrate contextual cues, with the descriptions in the above table suggesting that the model can overinterpret visual features, which is flagged as a limitation of the classifier, not least in relation to functional images (e.g., links or buttons). In line with the results in the previous section, these inconsistencies in the performance of the model seem to stem from inappropriate interpretation of contextual information in addition to the image itself, suggesting the need for a larger-scale training dataset. Such a dataset is in fact essential for generalisability, as the classifier remains highly sensitive to nuances in contextual cues in the training data and to the quality of the underlying annotations. The current classifier results therefore can only act as a proof-of-concept, validating the suggested approach of using a human-curated dataset for training models to tackle persistent alt text barriers; however, it is necessary that larger-scale similar training datasets are used to evaluate the potential of models to tackle such barriers at scale.

11.3 Context presence evaluation (S-RQ10)

Whilst the absence of context in training datasets has been highlighted as a key deficiency relating to the poor quality of alt text generated by SoTA V2L models (see section 4.4.1 and chapter 6), two variants of the second AI model that was developed in this thesis (see 9.2.3.3), which is Qwen2.5-VL-3B-Instruct, were trained on the GWAP-generated dataset (one variant was trained with and one without contextual information). The evaluation that followed relating to the semantic similarity of context prompts is similar to the procedure used in Section 10.3.3. The researcher investigated the binary presence of elements of the semantic definition of alt

text context (altC) (Section 4.3.1) in automatically generated descriptions. To achieve this, he investigated each alt text description generated by each variant of the model in tandem with its context prompt and assigned a binary value of one/zero to each element of the prompt that was present/absent in the alt text, respectively. Then, each alt text description received a **context presence score,** which was the total of all of its assigned binary values summed up, ranging from zero (no elements present) to five (all elements present). Finally, the average of all descriptions' presence score for each variant of the model were calculated and compared to help answer S-RQ10. The average context presence score of the variant trained on the dataset without contextual information was 1.64, while the variant with contextual information had an average presence score of 3.22. This highlights a significant improvement in the ability of the model to generate context-driven alt text descriptions; thus, validating the utility of structured context prompts for training V2L models.

Table 18. Distribution of context presence scores per model variant with improvement rates

Context presence score	No-Context (n)	With-Context (n)	Improvement (Δ n)	Improvement (Δ % points)
0	34	4	-30	-7.5
1	151	18	-133	-33.3
2	154	60	-94	-23.6
3	48	158	+110	+27.6
4	10	121	+111	+27.8
5	2	38	+36	+9.0
Total	399	399	N/A	N/A
Average score	1.64	3.22	N/A	N/A

The above table further adds to the utility of context prompts for the automated generation of more context-driven alt text descriptions by showing the distribution of scores across both variants of the model. The table shows that the variant of the model trained with contextual information achieved a reduction of 64 percentage points in alt text descriptions with low context presence scores (0-2) from 339 to 82 (-257 descriptions), while descriptions with high presence scores (3-5) also increased by 64 percentage points (+257 descriptions). These improvement rates are significant and explain the almost double average context presence score improvement (from 1.64 to 3.22) when learning from structured context prompts. It can thus be surmised that embedding contextual information (structured context prompts) within the training dataset can significantly help the model learn to generate more context-driven alt text.

11.4 Chapter summary

This chapter presented the evaluation of the performance of the AI models proposed in this thesis, i.e., HumanALT-O-matic and ContextALT-O-matic, that were fine-tuned and trained

on the GWAP-generated dataset compared to their control versions. First, the trained version of HumanALT-O-matic revealed a substantial improvement (75%) of automatically generated alt text (as per participants' quality perceptions) compared to the control version, following an online survey with 17 former TagALTlong players (Section 11.2). This was also investigated by examining the three strongest cases where the trained version was preferred over the control model and vice versa, as well as three illustrative non-statistically significant cases (Table 16). As a result, it was shown that the effectiveness of training is most evident when the quality of alt text generated by the control model is poor, with this improvement being linked to the ability of the trained version to correctly interpret contextual information. Second, the evaluation of ContextALT-O-matic through investigation of the binary presence of elements of the context definition (see section 4.3.1) revealed significant improvements in average presence score in the trained version (3.22/5) compared to the control version (1.64/5). This was corroborated by the distribution of context presence scores across both versions of the model, with the trained version achieving a reduction and an increase of 64 percentage points in low scores (0-2) and in high scores (3-5), respectively (Table 18). These findings underscore the validity of context prompts for training the AI model, which nearly doubles (from 1.64 to 3.22 average context presence score) its ability to generate context-driven alt text. Therefore, the work in this chapter addresses the lack of evaluation of the performance of AI models in terms of accessibility in relevant literature and highlights the effectiveness of using a human-curated output to train such models for automatically generating better, context-driven alt text.

Chapter 12. Findings and concluding discussion

12.1 Introduction

This chapter discusses the overall findings of the research work conducted in this thesis and presents the identified contributions and provides answers to the overarching RQs defined in Chapter 1. Finally, the chapter highlights limitations and avenues for future research work.

12.2 Overall findings

Following an abductive approach to theory development (Chapter 7), a scoping review of the web accessibility literature was conducted (Chapter 2) to understand the current state of web accessibility efforts and the holistic impact of web accessibility barriers on the wider range of disabilities. Past work has in fact produced similar reviews (Abdel-Wahab *et al.*, 2019; van der Smissen *et al.*, 2020; Acosta-Vargas *et al.*, 2021; Jonsson *et al.*, 2023), but they were largely restricted to specific sectors and disabilities, such as healthcare and audiovisual disabilities. Whilst existing efforts are most often limited to one type of disability (Friedman and Bryen, 2007; Brady *et al.*, 2013; Morris *et al.*, 2018; Muehlbradt and Kane, 2022), or are reporting the general prevalence of each barrier on the web (WebAIM, 2020; 2025), the review allowed for constructing and proposing a framework to assess the impact of barriers **across** the diversity of disabilities (Chapter 3). The framework responds to the lack of a quantifiable measure for the impact of each barrier across disabilities (Aizpurua, Harper and Vigo, 2016; Vollenwyder *et al.*, 2023), and it is the first attempt to measure the impact of barriers by taking into account disability-specific considerations. Deferring to the framework, alt text barriers, i.e., missing alt text and unsuitable alt text, were highlighted, with the latter being particularly underexplored, pointing towards the need to explore these barriers and current solutions (Chapter 4).

The review on alt text revealed lacklustre improvements in the case of suitability compared to alt text availability, while also noting the difficulty of the task and the reluctance to author alt text. Additionally, the central role of the **context** in which the image is used in was discussed and a relevant structured semantic definition of Alt Text Context (altC) was proposed (Section 4.3.1) to respond to this concept being loosely defined (Edwards *et al.*, 2023). This was shown to be particularly relevant for solutions to alt text barriers, as the increase in multimedia content on the Web prohibits manual alt text authorship, but automated solutions through AI fail in the area of alt text suitability, due to the lack of context in training data (Zong *et al.*, 2022). It was further shown that accessibility experts' takes on the suitability of alt text vary (Das *et al.*, 2024), pointing towards the need to investigate crowdsourcing solutions that hand out alt text authorship to large non-expert crowds. The complexity of authoring alt text suitably, however,

necessitates that such approaches incorporate training of non-experts (Miranda and Araujo, 2022), which led to the review of **GWAPs** (game-based crowdsourcing) (Chapter 5).

GWAPs have in fact been shown to be well-suited for incorporating training of non-experts (Tuite, 2014) and for large-scale data collection due to the lack of monetary incentives (Madge, 2020). However, there is a lack of GWAPs for the purpose of alt text authorship and evaluation. The approach has nonetheless shown promise for gathering large-scale datasets to train AI models for similarly complex tasks (e.g., protein folding (Curtis *et al.*, 2015), and linguistic annotation (Lafourcade and Lebrun, 2023)). This led to the need to explore the effectiveness of the approach for training AI models for automatically generating alt text to address the reluctance of authors and the increased abundance of multimedia content on the Web. The need to explore GWAP approaches was further highlighted by the poor quality of automatically generated alt text by AI models for image captioning (i.e., V2L captioning models), owing to their noisy web-scraped training datasets (Birhane, Prabhu and Kahembwe, 2021). The review of SoTA V2L models and their training datasets in fact revealed the lack of context in training data and the lack of evaluation of the performance of models on accessibility as key gaps in the current efforts (Chapter 6).

Accordingly, the need to automate alt text generation with the use of an alternative approach to large-scale data collection, due to the irrelevance of web scraping for accessibility, led to the proposal of a multi-faceted solution, comprising three infrastructures (GWAP, Cloud, and ML) (Chapter 9). It was, nonetheless, first necessary to determine actionable requirements based on the perspectives of alt text consumers and creators to guide the design of the proposed solution. Therefore, semi-structured interviews with both VIUs and WCCs were conducted (Chapter 8), corroborating a reported mismatch in the literature between their views on alt text suitability (Harris, 2020). Both participant groups, however, stressed the need for solutions to incorporate trainability of non-expert alt text authors to turn them into pseudo-experts, which agrees with the varying opinions of accessibility experts (Mack *et al.*, 2021); thus, further highlighting the relevance of proposing GWAPs to address alt text suitability. Past evidence on the desire of VIUs to become active alt text authors (Chisholm and Henry, 2005; Vollenwyder *et al.*, 2023) was also refuted, but collaborative efforts that position VIUs as evaluators of accessible content were encouraged. Qualitative insights from the interviews also produced a set of trainability (Sections 8.5.1.4 and 8.5.2.4) and alt text suitability recommendations (Table 7), which further highlighted the need for context-driven alt text and were used to guide the design of the GWAP developed in this thesis (TagALTlong).

Moreover, 125 players were recruited to play TagALTlong as part of a user study over a six-week period (Chapter 10), authoring 1208 and rating 1836 alt text descriptions. The analysis revealed a positive level of consistency among player ratings (Fig. 22) strengthening the potential of the GWAP solution, as it is well established that tasks like image caption rating can yield dependable results when aggregated over multiple workers (Simons *et al.*, 2020), which is the case in this work where participants acted as a crowd of evaluators. Further, the findings of the study aligned with past research (Kapur and Kreiss, 2024) emphasising that blind and low vision people often prefer longer, more informative image descriptions in contrast to sighted people. In the context of this study, verbose alt text descriptions were judged as higher quality by players, highlighting the potential of the GWAP approach to train non-expert annotators to align with the expectations of blind and low vision people. This further echoes earlier studies reporting that descriptions should go beyond minimal labels, especially in news or social media webpages (Stangl *et al.*, 2021). Verbosity-wise, TagALTlong players' highest rated alt text descriptions had an average character count of 140.2, aligning with past studies and accessibility documentation, which advise that descriptions rarely exceed 150 characters (W3C, 1999; Williams et al., 2022). At the same time, some studies highlight that more detailed descriptions are not always beneficial, warning that the type of details and the context of use are important considerations (Aguirre *et al.*, 2022). Thus, these findings should be interpreted in context rather than producing alt text descriptions to be always verbose.

A further finding of this user study was that context-driven alt text descriptions positively influence their quality, which is also supported by prior work. For instance, Gubbi Mohanbabu and Pavel (2024) showed that blind and low vision people rated context-driven image captions significantly higher than captions with no context. This is particularly important as it is known that most automatic captioning approaches typically ignore context (see section 4.4.1), which is suboptimal considering that the same description can fail on one website but be very useful on another depending on context (Miller, 2022). Moreover, the demonstrated potential of the proposed GWAP approach is consistent with the broader success of game-based crowdsourcing. It has in fact been shown that GWAPs can mobilise large numbers of volunteers for annotation tasks (e.g., see the ESP Game for image labelling (Von Ahn and Dabbish, 2004)), and while it can be argued that alt text annotation is a difficult task that also needs to boast trainability of non-expert annotators, more recent GWAPs were successfully used for difficult tasks (e.g., see PhraseDetectives for anaphora resolution (Yu et al., 2022), and JeuxDeMots for lexical relation extraction (Lafourcade and Le Brun, 2023)). The findings of the study echo these successes, as

by embedding annotation and evaluation in gameplay, many players were able to contribute to scalable alt text generation simultaneously.

Nevertheless, and although the current deployment demonstrates that scaling the generation of alt text through GWAPs is feasible, successfully doing so with the live release of a GWAP on the Web entails further challenges. Player retention, for example, would need to be taken into account, which is challenging even for some very successful GWAPs for tasks that are similarly complex with alt text annotation (Yu et al., 2022; Lafourcade and Le Brun, 2023). Although the participant pool gathered was very satisfying in the context of the present study numbers-wise, it is unclear whether player recruitment would be similar in a Web context, as there would be no participation call and different incentives would need to be considered. In effect, player motivation and retention are considered key challenges in the design of GWAPs that have yet to be addressed sufficiently (Chamberlain *et al.*, 2013, Droutsas, 2021). Thus, these challenges are different and need to be considered in addition to challenges addressed in this work when a GWAP for alt text annotation is pushed live on the Web, and they are particularly relevant considering that there is no reported GWAP for this purpose.

Whilst a standalone GWAP solution entails benefits for tackling alt text barriers, the need to automate alt text generation to scale such generation on par with modern needs and to address the reluctance to author alt text was addressed with the use of the GWAP-generated dataset to train AI models for automatically generating human-centred alt text (Chapter 11). In fact, two models were developed and used in this evaluation; i.e., the first model (HumanALT-O-matic), which was based on Google's BERT (for classification) and T5-small (for caption generation) sought to approximate average human-level quality alt text while automating alt text generation; and, the second model (ContextALT-O-matic), which was based on Qwen2.5-VL-3B-Instruct (Bai *et al.*, 2025) sought to automatically generate more context-driven alt text via training on context prompts (Section 4.3.1). There were two versions of each of these models, i.e., one version was fine-tuned and trained on the GWAP-generated datasets, while the second was a control version of the model that was not trained on the dataset. HumanALT-O-matic was evaluated via an online survey with a sample of 17 former players of TagALTlong, who rated alt text generated by both versions of the model for 20 unique pairs of image and context. Results served as a proof of concept, revealing a significant improvement in the quality of generated alt text by the trained version compared to pure image processing, while also showing the former's potential to classify decorative images. Prior work has in fact shown that there is a pressing need to automate this classification (Lengua, Rubano and Vitali, 2022) to address further barriers, such as the null alt text barrier (Caprette, 2025). ContextALT-O-matic

was evaluated via investigation of the binary presence of context definition (Section 4.3.1) elements in automatically generated alt text descriptions by the control and the trained version of the model. The investigation validated the effectiveness of the context prompts to help the model better learn how to generate more context-driven alt text descriptions, whose quality was shown both in the literature (Mack *et al.*, 2021) and in the findings of the second user study (Chapter 10).

Prior work on V2L captioning models has demonstrated the ability of models (e.g., PALI-17B, InternVL-Chat) to achieve SoTA results on captioning benchmarks (X. Chen *et al.*, 2023; Chen *et al.*, 2024), but there is a gap in the performance of these models relating to accessibility. The evaluation of HumanALT-O-matic and ContextALT-O-matic in terms of alt text quality and context presence address this gap by using a community-focused, human-centred approach to collect training data instead of web scraping, which has proven irrelevant for accessibility (Birhane, Prabhu and Kahembwe, 2021). Although efforts to incorporate context in training datasets are not new (Laurençon *et al*., 2023; Ramos *et al.*, 2023), reported efforts were also based on web-scraped data and were not evaluated on accessibility. However, the results serve as a proof-of-concept, validating the use of a human-curated dataset for training V2L models, but it is necessary that larger-scale training datasets are used to evaluate the potential of models to tackle alt text barriers at scale.

Taken together, the findings of this research commenced with exploring the landscape of web accessibility literature, revealing negligible improvements with regard to the suitability of alt text on the Web in recent years (Chapter 2). It was then further revealed that lack of context in alt text is a key deficiency in the low quality of alt text on the Web. The altC definition of context in alt text was thus developed, with the lack of context in training datasets being also highlighted as a pivotal reason for the poor quality of alt text that has been automatically generated by ML models (Chapter 3). It was also shown that WCCs, who are responsible for providing alt text are reluctant to both learn how to author and to author alt text for web content, due to task tediousness, while accessibility experts' views on what makes alt text suitable vary gravely (Chapter 4). These gaps then pointed towards GWAPs, as a crowdsourcing approach that involves non-experts to address variable expert views, as well as being able to train players using the altC definition of context, which is essential for alt text suitability (Chapter 5). The GWAP, thus, generated the community-sourced, context-driven dataset that was used to fine-tune two proof-of-concept ML models (Chapter 9). This was a necessary step to complete the proposed system solution, owing to the surge of AI technologies and the increased multimedia abundance on the web, in tandem with the need for alt text to be available and suitable for all

web images. Therefore, the proposed solution is unique and novel, as it actioned through a GWAP approach, whose design and development are anchored in web accessibility theory (incl. gaps and barriers), with the resulting human-curated, context-driven data being used to train ML models, as opposed to web-scraped data, which have no implications for accessibility.

12.3 Contributions and research objectives

This work and findings link to the research objectives (OBJ1-OB7) designated at the beginning of this thesis (see section 1.3). Accordingly, this research has resulted in the appreciation of a plurality of perspectives on web accessibility, including those of scholarly work (OBJ1), web content creators and consumers (OBJ3), allowing for the development of a holistic view on the impact of accessibility barriers across diverse disabilities. Alt text barriers were highlighted and explored through various angles in this research, including survey work to find gaps in the suitability of alt text (OBJ2), interviews with VIUs and WCCs to understand the perception mismatch between them (OBJ3), and the development of practical solutions to address alt text barriers (i.e., the proposed GWAP) (OBJ4). Several key needs for tackling alt text barriers were highlighted in this research; i.e., not relying on conformance with accessibility guidelines, including context in alt text, involving and training non-experts in alt text authorship, and automating alt text generation without compromising quality. These were identified through a multi-staged research approach (Chapter 7), resulting in the development of a practical GWAP solution to tackle alt text barriers and help improve automated alt text generation through AI models (OBJ4). Overall, the main contributions of this thesis therefore are as follows:

- An Impact Assessment Framework, which is the first, to the best of the researcher's knowledge, attempt to measure the impact of web accessibility barriers by taking into account disability-specific considerations (OBJ1) (Chapter 3).
- The first structured semantic definition and syntax of **Alt Text Context (altC)**, accounting for multiple factors that influence how an image should be described in alt text and which is framed within two important elements related to the image and the webpage being used in. It can serve as a framework towards improving the relevance and informativeness of alt text beyond traditional isolated image labelling (Chapter 4).
- Alt text trainability and suitability recommendations drawing on a set of six themes, which are to the best of the researcher's knowledge, one of the first such efforts to compare and bring together the views of web content creators and visually impaired users on alt text suitability (OBJ3) (Chapter 8).

- The first, to the best of the researcher's knowledge, GWAP for alt text annotation and evaluation that embeds context by design, which is also one of the very few crowdsourcing efforts for such type of annotation (OBJ4) (Chapter 9).
- A novel dataset of player-authored alt text descriptions and rating scores that considers context in alt text, which is available upon reasonable request[15] (OBJ5) (Chapter 10).
- Two proof-of-concept AI models that integrate contextual features during training and assessing their impact in shaping model performance, thereby offering empirical evidence for the impact of contextual cues on alt text generation (Chapter 11).

12.4 Research questions revisited

Deferring to these contributions, the overarching research questions (RQs) of this thesis (see chapter 1) can be answered as follows:

RQ1. What are the main web accessibility barriers and what is their impact across different user groups of people with disabilities?

Through the review in Chapter 2, web content was found to pose considerable challenges for both content creators and consumers. On the one hand, the proliferation of new technologies (e.g., Virtual Reality and Mixed Reality, Artificial Intelligence, etc.) is allowing consumers to interact with web content in new, more engaging, and fun ways, but on the other hand the widely reported issues with missing and poor-quality alt text, and low contrast text even within the most fundamental web content demonstrate that a solution for these pivotal accessibility barriers has yet to be found. This work further identified that there are currently no efforts to the best of the researcher's knowledge to determine the impact of each barrier across diverse disabilities; hence, a much-needed related quantifiable measure was proposed and presented in the form of CxDAI (see section 3.2).

RQ2. What are recommendations for practice based on the perspectives of both visually impaired users and web content creators on alt text suitability?

RQ2 was addressed by first addressing the below S-RQs through an empirical user study with both VIUs and WCCs (Chapter 8):

[15] The dataset is available from the researcher - Email: nick.droutsas@brunel.ac.uk

- S-RQ1. What are the perceptions of web content creators on the accessibility of the web through screen readers against visually impaired users' web navigation experiences?
- S-RQ2. What are the perceptions of web content creators on WCAG against those of visually impaired users?
- S-RQ3. What makes alt text suitable according to both visually impaired users and web content creators?

Through the user study with VIUs and WCCs, the latter's lack of experiential understanding of web navigation via screen readers was highlighted. Both groups essentialised the ironclad role of training in alt text authorship, which is primarily the responsibility of web content creators, but collaborative efforts with visually impaired users as alt text evaluators were acknowledged. In response, a set of trainability recommendations (Sections 8.4.1.4 and 8.4.2.4) were presented, specifying the need for training on suitability to be structured and example-driven. A set of recommendations for alt text suitability was also proposed (Table 7), which are to the best of the researcher's knowledge the first such recommendations that bring together the views of both VIUs and WCCs.

RQ3. Is a GWAP an efficient approach to the generation of human-centred, context-driven alt text at scale?

RQ3 was addressed by first addressing the below S-RQs through an evaluation of the player-authored and rated alt text descriptions in the TagALTlong GWAP (Chapter 10):

- S-RQ4. How effective is the implemented solution in generating alt text descriptions at scale?
- S-RQ5. How does the use of structured context prompts influence the quality of player-authored alt text?
- S-RQ6. How does the descriptive verbosity of player-authored alt text influence its quality perception?
- S-RQ7. What levels of consistency or divergence emerge among player ratings?

One hundred and twenty-five participants with no prior expertise in alt text were recruited to play the TagALTlong GWAP (Chapter 9), as part of an empirical user study over a six-week period. A human-centred dataset of 1208 authored and 1836 rated alt text descriptions was gathered, highlighting GWAPs as a promising approach to address alt text generation at scale, as more than 1000 alt text were collected in a small timeframe. The mixed-method analysis of

results based on crowdsourced annotations, user ratings and consensus measures showed that more verbose and context-rich alt text descriptions are rated higher in quality. Taken together, these results highlight GWAPs as a practical scalable approach to enhance alt text generation for accessibility.

RQ4. Is it possible to generate human-level quality alt text descriptions whilst automating alt text generation through AI?

RQ4 is the final RQ of this thesis, and it can be answered thanks to the completion of all user studies (Chapters 8 and 10), solution development (Chapter 10), and evaluations (Chapter 11). It was also necessary to first address the below S-RQs through performance evaluation of the AI models trained on the GWAP-generated dataset compared to pure image processing:

- S-RQ8: Is the use of a GWAP-generated dataset to train the AI model (trained model) for generating alt text descriptions an effective approach to get closer to a human average compared to pure image processing (control model)?
- S-RQ9: How well does the trained AI model classify decorative (eye candy) images?
- S-RQ10: To what extent does learning from structured context prompts improve context-driven alt text generation?

The analysis of the online survey results relating to the quality of alt text generated by the AI models act as a proof-of-concept, validating the use of a human-curated dataset deriving from a GWAP for training models to tackle alt text barriers. The ability of the trained version of the model (HumanALT-O-matic) to classify not only the image itself but also its role in the context it is used in was shown, whilst it is vital to gather larger-scale similar training datasets to address alt text barriers at scale. It was also revealed that the version of ContextALT-O-matic that was trained on structured context prompts saw a significant improvement in its ability to automatically generate more context-driven alt text compared to the control version.

12.5 Research limitations

The research and findings in this thesis present some limitations that need to be considered. First, it is acknowledged that the review on web accessibility and barriers is limited to the results of the search (Chapters 2 and 3). Second, although every effort was made to capture web accessibility preferences, activity and barriers across diverse disabilities, it is disclosed that the findings may represent an unbalanced representation of research work per disability.

Third, it is acknowledged that the CxDAI measures (Table 1) derive from different studies and that the taxonomy is disability-specific. Deferring to the findings of the first user study (Chapter 8), it is first acknowledged that the sample size does not allow for wide generalisations to be drawn from the reported conclusions (Section 8.5). In line with past research (Muehlbradt and Kane, 2022), the characteristics of the target sample made it difficult to recruit a larger sample of participants; however, it has to be noted that repetition was observed in the participants' answers after eight participants, for both groups.

Whilst it is recognised that richer qualitative data would have benefited this work, the sample population size is in line with peer studies (e.g., Mack *et al.*, 2021; Lee and Ashok, 2022), and the analysis maximises the data obtained therein. To the best of the researcher's knowledge, the sample is unique in comparing the views of VIUs and WCCs in the same study. Further, the WCC participants had an average of nine years of web accessibility experience (Table 5), which is important to take into consideration when comparing the findings with past studies, as expertise in web accessibility has been shown to vary. Importantly, it is clarified that the findings in relation to the WCC participants reflect a blend of their personal experience in creating web content and their observations of their clients in doing so. It is therefore recognised that contradictions with past evidence can very well be due to past evidence being based on data from WCCs that were less experienced in web accessibility. Whilst diverse impairments were acknowledged in the context of alt text barriers, this work only focused on such barriers as experienced by VIUs.

Deferring to the findings of the second user study (Chapter 10), it is most important to first consider that many images received fewer than ten alt text ratings; thus, alternatives to robust traditional inter-rater reliability (IRR) measures were used. Indicatively, a semantic similarity analysis was applied to investigate whether the contextual prompt—grounded in the proposed definition of alt text context—led to a good level of consistency in how participants rated the same description (Section 10.3.5). Despite this limitation, the results support the conclusion that the GWAP solution can provide a scalable mechanism for rapidly generating alt text, even if rater agreement is imperfect. Whilst it is recognised that these measures do not replace IRR, they provide an alternate lens through which to assess the coherence of user-generated content under a shared prompt. The dataset also excluded descriptions with none or only one rating in order to ensure minimal reliability in agreement measures; while this improved consistency, it reduced dataset size and may have excluded potentially valuable edge cases.

Whilst the tiered analysis strategy presented in this work (≥2 ratings for general reliability; ≥3 ratings for outlier detection) could inform data quality protocols in crowdsourced alt text

annotation tasks, the reliance on a single moderator for certain subjective ratings (e.g., presence score), which while methodologically justified, it may introduce some bias and may reflect individual interpretation. While moderation was only used when deemed as appropriate, the absence of a second independent moderator limits the ability to generalise some findings. This limitation is shared in the evaluation of the performance ContextALT-O-matic (Section 11.3), where the binary presence of context definition (Section 4.3.1) elements was investigated by a single moderator in alt text generated by the trained and the control version of the model. It is also recognised that non-parametric statistical tests were used for the performance evaluation of HumanALT-O-matic, as the data was not normally distributed. Whilst the discussed protocol and sampling strategy (Section 11.2.1) achieved the selection of a sample representative of the population, it is recognised that parametric tests would have strengthened the findings.

Overall, the most critical limitation of the current approach for real-world deployment is the discussed reliance on a single moderator to assess disagreement among raters in the GWAP and contextual representation in the alt text generated by the ContextALT-O-matic model. This creates a major bottleneck in a production context, as it introduces potential bias, interfering with both the validity and the reliability of outputs. Mitigation measures were in line, and as per the above discussion, involving the use of a moderator rating score only in extreme cases of disagreement among raters, and the mere check for the binary presence of elements of the altC definition, rather than assigning additional rating scores.

12.6 Future work

The research work carried out in this thesis also present several avenues for future research to explore. First, similar surveys of the landscape of web accessibility literature that are inclusive of a broader range of venues and application areas to identify additional work in accessibility research are recommended. Accordingly, more studies reporting on the preferences, perceived barriers, web activity and barrier disturbance rates of people with diverse disabilities are pivotal future steps to achieve more accurate measurability of the impact of accessibility barriers across disabilities. Similarly, a multinational study is needed to enhance the CxDAI framework by providing a more comprehensive understanding of global barriers and facilitating cross-country comparisons. Research efforts investigating the applicability of the CxDAI framework beyond the Web (incl. mobile apps, virtual environments, and other emerging technologies) are very much encouraged. A further need for future work lies in carrying out qualitative work similar to the interviews conducted with VIUs and WCCs (Chapter 9), as although the current findings agree with recent similar findings on the need for training in alt text authorship (WebAIM,

2025), they also contradict past work on the need for VIUs to become active alt text authors (Heylighen *et al.*, 2017; Vollenwyder *et al.*, 2023).

In a similar vein, more studies comparing the views of WCCs and VIUs in relation to web navigation and alt text suitability are much needed, and it is recommended that such studies disclose the types of disabilities, web accessibility experience of participants, and assistive technologies used. Accordingly, studies that compare the views of novice and expert WCCs are also encouraged to inquire into the validity of the findings of this study. Relatedly, mixed-group focus studies that belong to both participant groups (e.g., VIU participants who are also WCCs) are recommended as a valuable future endeavor to further explore contrasting views between the two groups and help identify pathways towards reconciliation. Further similar studies focusing on diverse impairments are also encouraged to capture the broader scope of alt text barriers and inform more inclusive practices.

Deferring to the analysis of GWAP-generated data, future research should apply more robust IRR metrics to investigate evaluator agreement, which was not achieved in this cause due to the aforementioned dataset constraints. The latter could be addressed in future work by aiming to increase the number of ratings per image by recruiting a larger pool of participants or by designing the GWAP for redundancy (e.g., dynamically routing more raters to underrated items). Further, incorporating stratified sampling or incentives for rating could help ensure a more even distribution of ratings across the dataset, thereby improving the reliability and validity of agreement metrics. On that note, additional analysis of gender-based differences with regard to alt text quality should also be considered to explore potential gender-driven differences. In this regard, the development of semi-automated moderation systems that can flag description with high rating variance or lexical anomalies for review, is another promising avenue for future work. Utilising natural language processing (NLP) techniques could also help assess semantic features (e.g., object coverage, sentiment or clarity) for assessing descriptive richness and quality in more meaningful ways and are thus worth exploring.

Whilst the results relating to the performance of the trained versions of the AI models used in this work showed improvements compared to pure image processing, the effectiveness of the GWAP-generated dataset for training such models was validated. Future work can use this to develop tools, such as arbiters or discriminators, for context representation accuracy, and could then provide a rating score with which to benchmark the quality of alt text that people write. It will therefore be made possible to automatically assess the quality of alt text based on the image and context. It is also recommended that similar evaluations are based on normally

distributed data to enable the use of parametric statistical tests to further highlight possibilities to approximate average human-level alt text quality whilst automating alt text generation.

Overall, it is the formal expansion, further operationalization and empirical validation of the altC definition of context in alt text that offers the strongest **theoretical** contribution in this research. As previously detailed, context in alt text is considered essential by recent scholarly work (McCall and Chagnon, 2022; Muehlbradt and Kane, 2022), and guidelines (BBC, 2023a, 2023b), yet remains loosely defined. This is also evident in the poor performance of state-of-the-art V2L models that are instead trained on web-scraped data for scalability, with several studies reporting on their irrelevance for screen reader use due to the lack of context in their training datasets (e.g., Birhane, Prabhu and Kahembwe, 2021; Desai *et al.*, 2021). Expanding on the definition and template proposed in this study with the altC is very much recommended avenue for future research seeking to advance thinking in accessibility, with implications both with regard to deepening the discourse on what makes alt text suitable and the use and role of context in the broader automated pipeline for the generation of human-level alt text.

12.7 Thesis conclusion

This thesis has shown the potential of using a game-based crowdsourcing approach aimed at alt text barriers, not least in relation to alt text suitability and context. It has highlighted the impact of context in alt text and presented a relevant novel definition (Section 4.3.1), which can be semantically adapted for incorporation into future work. The importance of context in alt text and contrasting perspectives of both WCCs and VIUs were appreciated, shedding light into a the largely unsharpened area of alt text suitability. It is the hope of the researcher that by disclosing these empirical perspectives in tandem with the promising results of the GWAP—TagALTlong—more GWAPs for context-driven alt text annotation will be developed, drawing on recommendations from WCCs and VIUs. The strength of crowdsourcing being in gathering datasets to automate hitherto manual tasks, TagALTlong was effective in generating a human-curated dataset (alt text descriptions and rating scores) of sufficient scale, considering the timeframe, to train AI models. Additionally, the models can be scaled to other domains (e.g., different web or digital contexts, graphic types, and usage environments) based on how far the contextual factors in the altC framework can be extrapolated to map to such domains' requirements. Singh *et al.* (2024), for example, showed that alt text for graphics in academic publications requires different considerations in alt text, which in turn needs an adaptation of the altC to guide the collection of training data for the models. Scalability can in this sense also be affected by differences in the visual information in graphics (e.g., from common objects to

specialised diagrams), and semantic interpretations of such content in different environments (e.g., organisational or technical environments). It can thus be surmised that the more palatable such considerations are to align with the characteristics of the training data for a given domain, the more minimal the fine-tuning that will be required for the model to scale to said domain. Whereas the current model pipelines are thus scalable to other domains in principle, it is noted that further data collection through GWAP gameplay is required when scaling to domains, such as Virtual Reality (VR) or Augmented Reality (AR), where combinations of visual, contextual, and functional factors are blurred beyond their counterparts on the Web. Data collection would, in such cases, also need to be domain-specific for effective adaptation.

As previously discussed, the GWAP approach is also effective in incorporating training of non-experts to turn them into pseudo-experts in alt text authorship, which is a key missing piece in current solutions. The positive effect of in-game training of non-experts, and the improved performance of models trained on the dataset in terms of alt text quality and context presence were highlighted as key claims to fame of the GWAP approach for tackling alt text barriers. This work presented a novel proof-of-concept in the form of —TagALTlong — the first GWAP for context-driven alt text annotation and evaluation, which showed promise as a standalone crowdsourcing game-based solution and as a foundation for automating context-driven alt text generation through AI.

Reference list

Abascal, J. and Nicolle, C. (2005) 'Moving towards inclusive design guidelines for socially and ethically aware HCI', *Interacting with computers*, 17(5), pp. 484–505.

Abdel-Wahab, N., Rai, D., Siddhanamatha, H., Dodeja, A., Suarez-Almazor, M.E. and Lopez-Olivo, M.A. (2019) 'A comprehensive scoping review to identify standards for the development of health information resources on the internet', *PLoS ONE*, 14(6), e0218342. Available at: https://doi.org/10.1371/journal.pone.0218342

AbilityNet (2017) *Free Digital Accessibility Resources | AbilityNet*. Available at: https://abilitynet.org.uk/accessibility-services/useful-resources (Accessed: 11 December 2023).

Abou-Zahra, S., Brewer, J. and Cooper, M. (2018) 'Artificial intelligence (AI) for web accessibility: Is conformance evaluation a way forward?', in *Proceedings of the 15th International Web for All Conference*, pp. 1–4.

Abuaddous, H., Zalisham, M. and Basir, N. (2016) 'Web accessibility challenges', *International Journal of Advanced Computer Science and Applications*, 7(10), pp. 172–181. Available at: https://doi.org/10.14569/IJACSA.2016.071023.

Accessibility for Ontarians with Disabilities Act (2023) 'Image Descriptions in Websites and Documents', *Accessibility for Ontarians with Disabilities Act (AODA)*, 22 November. Available at: https://aoda.ca/image-descriptions-in-websites-and-documents/ (Accessed: 9 January 2025).

Acosta-Vargas, P., Salvador-Acosta, B., Salvador-Ullauri, L., Villegas-Ch, W. and Gonzalez, M. (2021) 'Accessibility in native mobile applications for users with disabilities: A scoping review', *Applied Sciences*, 11(12), p. 5707.

Adhabi, E. and Anozie, C., 2017. 'Literature review for the type of interview in qualitative research'. *International Journal of Education*, 9(3), pp.88–91. Available at: https://doi.org/10.5296/ije.v9i3.11483

Aguirre, C.A., Mahmood, A. and Huang, C.-M. (2022) 'Crowdsourcing Thumbnail Captions via Time-Constrained Methods', in *27th International Conference on Intelligent User Interfaces. IUI '22: 27th International Conference on Intelligent User Interfaces*, Helsinki Finland: ACM, pp. 36–48. Available at: https://doi.org/10.1109/CoG60054.2024.10645588.

Aizpurua, A., Harper, S. and Vigo, M. (2016) 'Exploring the relationship between web accessibility and user experience', *International Journal of Human-Computer Studies*, 91, pp. 13–23.

Aliady, W. and Poesio, M. (2024) 'Master the Linguistic Landscape: Puzzle Integration in a

3D NLP Game', in *2024 IEEE Conference on Games (CoG)*. IEEE, pp. 1–8. Available at: https://ieeexplore.ieee.org/abstract/document/10645588/ (Accessed: 23 February 2025).

Aliady, W.A., Aloraini, A., Madge, C., Yu, J., Bartle, R. and Poesio, M. (2022) 'Coreference annotation of an arabic corpus using a virtual world game', in *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pp. 388–393. Available at: https://aclanthology.org/2022.wanlp-1.37/ (Accessed: 23 February 2025).

Amado-Salvatierra, H.R., Hilera, J.R., Tortosa, S.O., Rizzardini, R.H. and Piedra, N. (2016) 'Towards a semantic definition of a framework to implement accessible e-learning projects', *Journal of Universal Computer Science*, 22(7), pp. 921–942.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J. and Zhong, H. (2025) *Qwen2.5-VL technical report*. arXiv preprint arXiv:2502.13923. Available at: https://doi.org/10.48550/arXiv.2502.13923 (Accessed: 21 September 2025).

Bateman, D.R., Brady, E., Wilkerson, D., Yi, E.H., Karanam, Y. and Callahan, C.M. (2017) 'Comparing crowdsourcing and friendsourcing: a social media-based feasibility study to support Alzheimer disease caregivers', *JMIR research protocols*, 6(4), p. e6904.

BBC (2023a) *Guides - Accessibility for Products - BBC*. Available at: https://www.bbc.co.uk/accessibility/forproducts/guides/ (Accessed: 5 May 2024).

BBC (2023b) *Image alternatives - Accessibility for Products - BBC*. Available at: https://www.bbc.co.uk/accessibility/forproducts/guides/html/image-alternatives/ (Accessed: 5 May 2024).

BBC (2023c) *Mobile Accessibility Guidelines - Accessibility for Products - BBC*. Available at: https://www.bbc.co.uk/accessibility/forproducts/guides/mobile/ (Accessed: 5 May 2024).

Bellscheidt, S., Metcalf, H., Pham, D. and Elglaly, Y.N. (2023) 'Building the Habit of Authoring Alt Text: Design for Making a Change', in *The 25th International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS '23: The 25th International ACM SIGACCESS Conference on Computers and Accessibility*, New York NY USA: ACM, pp. 1–5. Available at: https://doi.org/10.1145/3597638.3614495.

Bennett, C.L., Gleason, C., Scheuerman, M.K., Bigham, J.P., Guo, A. and To, A. (2021) '"It's Complicated": Negotiating Accessibility and (Mis) Representation in Image Descriptions of Race, Gender, and Disability', in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Bennett, C.L. and Rosner, D.K. (2019) 'The Promise of Empathy: Design, Disability, and Knowing the "Other"', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems*,

Glasgow Scotland Uk: ACM, pp. 1–13. Available at: https://doi.org/10.1145/3290605.3300528.

Berger, A., Caspers, T., Croll, J., Hofmann, J., Kubicek, H., Peter, U., Ruth-Janneck, D. and Trump, T. (2010) *Web 2.0/barrierefrei. Eine Studie zur Nutzung von Web 2.0 Anwendungen durch Menschen mit Behinderung*. Bonn: Aktion Mensch. Available at: https://medien.aktion-mensch.de/publikationen/barrierefrei/Studie_Web_2.0.pdf.

Bernard, R. (2019) *Web accessibility and mental disorders: difficulties experienced by people with depression and anxiety on the Web*. PhD Thesis. Ludwig-Maximilians-Universität. Available at: https://core.ac.uk/download/pdf/323446571.pdf (Accessed: 27 October 2023).

Bi, T., Xia, X., Lo, D., Grundy, J., Zimmermann, T. and Ford, D. (2022) 'Accessibility in software practice: a practitioner's perspective', *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(4), pp. 1–26.

Bianchi, T. (2023) *Colombia: number of internet users 2028*, Statista. Available at: https://www.statista.com/forecasts/1143947/internet-users-in-colombia (Accessed: 3 November 2023).

Birhane, A., Prabhu, V.U. and Kahembwe, E. (2021) *Multimodal datasets: misogyny, pornography, and malignant stereotypes*. arXiv. Available at: https://doi.org/10.48550/arXiv.2110.01963.

Brady, E., Morris, M. and Bigham, J. (2014) 'Friendsourcing for the greater good: Perceptions of social microvolunteering', in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pp. 6–7.

Brady, E.L., Zhong, Y., Morris, M.R. and Bigham, J.P. (2013) 'Investigating the appropriateness of social network question asking as a resource for blind users', in *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 1225–1236.

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77–101. Available at: https://doi.org/10.1191/1478088706qp063oa.

Braun, V. and Clarke, V. (2016) *Thematic analysis frequently asked questions*. Auckland: The University of Auckland. Available at: http://www.psych.auckland.ac.nz/en/about/our-research/research-groups/thematic-analysis/frequently-asked-questions-8.html

Braun, V. and Clarke, V. (2019) 'Reflecting on reflexive thematic analysis', *Qualitative Research in Sport, Exercise and Health*, 11(4), pp. 589–597. Available at: https://doi.org/10.1080/2159676X.2019.1628806.

Braun, V. and Clarke, V. (2021) 'One size fits all? What counts as quality practice in (reflexive) thematic analysis?', *Qualitative Research in Psychology*, 18(3), pp. 328–352. Available at:

https://doi.org/10.1080/14780887.2020.1769238.

Brewer, J. (2004) 'Web accessibility highlights and trends', in *Proceedings of the 2004 international cross-disciplinary workshop on Web accessibility (W4A)*, pp. 51–55.

Brewer, J. (2011) 'Accessibility of the World Wide Web: Technical and Policy Perspectives', in *Universal Design Handbook*. United States: McGraw-Hill.

Bureau of Internet Accessibility (2018) *Why Is It Important for Accessibility to Use Actual Text Instead of Images of Text?* Available at: https://www.boia.org/blog/why-is-it-important-for-accessibility-to-use-actual-text-instead-of-images-of-text (Accessed: 9 January 2025).

Burkett, R. (2013) 'An alternative framework for agent recruitment: From MICE to RASCLS', *Studies in Intelligence*, 57(1), pp. 7–17.

Byrne, D. (2022) 'A worked example of Braun and Clarke's approach to reflexive thematic analysis', *Quality & Quantity*, 56(3), pp. 1391–1412. Available at: https://doi.org/10.1007/s11135-021-01182-y.

Calás, M.B. and Smircich, L. (eds) (2018) *Postmodern management theory*. Abingdon: Routledge.

Campoverde-Molina, M., Luján-Mora, S. and Valverde, L. (2021) 'Process model for continuous testing of web accessibility', *IEEE Access*, 9, pp. 139576–139593.

Caprette, H. (2025) *Creating an Empty or Null Alt Attribute for Decorative Images – Best Practices in Accessible Online Design*. Available at: https://pressbooks.ulib.csuohio.edu/accessibility/chapter/chapter-3-6-creating-an-empty-or-null-alt-attribute-for-decorative-images/ (Accessed: 9 January 2025).

Carlsson, G., Iwarsson, S. and Ståhl, A. (2002) 'The Personal Component of Accessibility at Group Level: Exploring the Complexity of Functional Capacity', *Scandinavian Journal of Occupational Therapy*, 9(3), pp. 100–108. Available at: https://doi.org/10.1080/11038120260246932.

Chafetz, J. (2005) 'It's the Aggregation, Stupid (reviewing James Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations (2004))', *HeinOnline*. Available at: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/yalpr23&section=29 (Accessed: 6 January 2024).

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M. and Poesio, M. (2013) 'Using games to create language resources: Successes and limitations of the approach', in *The People's Web Meets NLP*. Springer, pp. 3–44.

Changpinyo, S., Sharma, P., Ding, N. and Soricut, R. (2021) 'Conceptual 12m: Pushing web-

scale image-text pre-training to recognize long-tail visual concepts', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568. Available at: http://openaccess.thecvf.com/content/CVPR2021/html/Changpinyo_Conceptual_12M_Pushing_Web-Scale_Image-Text_Pre-Training_To_Recognize_Long-Tail_Visual_CVPR_2021_paper.html (Accessed: 21 February 2025).

Chatziemmanouil, T. and Katsanos, C. (2024) 'Accessibility Academy: Interactive Learning of the WCAG 2.1 Web Accessibility Guidelines', in *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, pp. 1–7. Available at: https://ieeexplore.ieee.org/abstract/document/10578915/ (Accessed: 17 March 2025).

Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W. and Cohen, W.W. (2023) 'Subject-driven text-to-image generation via apprenticeship learning', *Advances in Neural Information Processing Systems*, 36, pp. 30286–30305.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L. and Kolesnikov, A. (2023) *PaLI: A Jointly-Scaled Multilingual Language-Image Model*. arXiv. Available at: https://doi.org/10.48550/arXiv.2209.06794.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L. and Li, B. (2024) 'Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198. Available at: http://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html (Accessed: 4 April 2025).

Chintalapati, S.S., Bragg, J. and Wang, L.L. (2022) 'A Dataset of Alt Texts from HCI Publications: Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations in Scientific Papers', in *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–12.

Chisholm, W.A. and Henry, S.L. (2005) 'Interdependent components of web accessibility', in *Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility (W4A)*, pp. 31–37.

Chrons, O. and Sundell, S. (2011) 'Digitalkoot: Making old archives accessible using crowdsourcing', in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Citeseer.

Cionca, E. and Kohler, T. (2024) *Alt Text: What to Write*, Nielsen Norman Group. Available at: https://www.nngroup.com/articles/write-alt-text/ (Accessed: 3 March 2025).

Clark, J. (2001) *Building accessible websites*. London: Pearson Education.

Clarke, V. and Braun, V. (2013) *Successful qualitative research: A practical guide for beginners*. Available at: https://www.torrossa.com/gs/resourceProxy?an=5017629&publisher=FZ7200 (Accessed: 9 May 2024).

Cooper, M. (2016) 'Web accessibility guidelines for the 2020s', in *Proceedings of the 13th International Web for All Conference*, pp. 1–4.

Cooper, S. (2022) 'A Case Study of Quality-Diversity Search in Human Computation', *Human Computation*, 9(1), pp. 58–65.

Crespo, R.G., Espada, J.P. and Burgos, D. (2016) 'Social4all: Definition of specific adaptations in Web applications to improve accessibility', *Computer Standards & Interfaces*, 48, pp. 1–9.

Creswell, J.W. and Poth, C.N., 2018. *Qualitative inquiry and research design: Choosing among five approaches*. 4th ed. Thousand Oaks, CA: Sage Publications.

Crow, L. (2010) 'Including all of our lives: Renewing the social model of disability', in *Equality, Participation and Inclusion 1*. Routledge, pp. 136–152.

Curtis, V. (2015) 'Motivation to participate in an online citizen science game: A study of Foldit', *Science Communication*, 37(6), pp. 723–746.

Das, M., Fiannaca, A.J., Morris, M.R., Kane, S.K. and Bennett, C.L. (2024) 'From Provenance to Aberrations: Image Creator and Screen Reader User Perspectives on Alt Text for AI-Generated Images', in *Proceedings of the CHI Conference on Human Factors in Computing Systems. CHI '24: CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, pp. 1–21. Available at: https://doi.org/10.1145/3613904.3642325.

Dellermann, D., Lipusch, N., Ebel, P. and Leimeister, J.M. (2020) 'The potential of collective intelligence and crowdsourcing for opportunity creation', *International Journal of Entrepreneurial Venturing*, 12(2), p. 183. Available at: https://doi.org/10.1504/IJEV.2020.105569.

Desai, K., Kaul, G., Aysola, Z. and Johnson, J. (2021) *RedCaps: web-curated image-text data created by the people, for the people*. arXiv. Available at: https://doi.org/10.48550/arXiv.2111.11431.

Dobransky, K. and Hargittai, E. (2016) 'Unrealized potential: Exploring the digital disability divide', *Poetics*, 58, pp. 18–28.

Dortheimer, J. (2022) 'Collective intelligence in design crowdsourcing', *Mathematics*, 10(4),

p. 539.

Droutsas, N. (2021) Gamers with the purpose of language resource acquisition: Personas and scenarios for the players of language resourcing games-with-a-purpose. Uppsala: Uppsala University. Available at: https://uu.diva-portal.org/smash/record.jsf?pid=diva2:1567059

Droutsas, N., Spyridonis, F., Daylamani-Zad, D. and Ghinea, G. (2025a) 'Flower in the Mirror, Moon on the Water: Bridging Perspectives on Alternative Text and Recommendations for Practice', *International Journal of Human–Computer Interaction*, pp. 1–21. Available at: https://doi.org/10.1080/10447318.2025.2499659.

Droutsas, N., Spyridonis, F., Daylamani-Zad, D. and Ghinea, G. (2025b) 'Web Accessibility Barriers and their Cross-disability Impact in eSystems: A Scoping Review', *Computer Standards & Interfaces*, 92, p. 103923. Available at: https://doi.org/10.1016/j.csi.2024.103923.

Edwards, E.J., Gilbert, M., Blank, E. and Branham, S.M. (2023) 'How the alt text gets made: What roles and processes of alt text creation can teach us about inclusive imagery', *ACM Transactions on Accessible Computing (TACCESS)*, 16, pp. 1–28. Available at: https://doi.org/10.1145/3587469

EEID, T.E.S.D. 2004 (2004) 'Adopted on 9 May 2004, at the Annual General Meeting of the European Institute for Design and Disability in Stockholm', *Design for All Europe*. Available at: https://dfaeurope.eu/what-is-dfa/dfa-documents/the-eidd-stockholm-declaration-2004/ (Accessed: 14 February 2023).

European Parliament and Council (2019) *Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services (Text with EEA relevance)*, OJ L. Available at: http://data.europa.eu/eli/dir/2019/882/oj/eng (Accessed: 11 December 2023).

Faulkner, L. (2003) 'Beyond the five-user assumption: Benefits of increased sample sizes in usability testing', *Behavior Research Methods, Instruments, & Computers*, 35(3), pp. 379–383. Available at: https://doi.org/10.3758/bf03195514

Fort, K., Guillaume, B. and Chastant, H. (2014) 'Creating Zombilingo , a game with a purpose for dependency syntax annotation', in *Proceedings of the First International Workshop on Gamification for Information Retrieval - GamifIR '14*. the First International Workshop, Amsterdam, The Netherlands: ACM Press, pp. 2–6. Available at: https://doi.org/10.1145/2594776.2594777.

Friedman, M.G. and Bryen, D.N. (2007) 'Web accessibility design recommendations for people with cognitive disabilities', *Technology and disability*, 19(4), pp. 205–212.

Games and NLP (2019) *Games and NLP 3 A Gamer's Perspective on GWAPs*. Available at:

https://www.youtube.com/watch?v=-Z2ba0ltSdk (Accessed: 9 August 2023).

Games and NLP (2020) *Games and NLP 2020 Keynote: Designing Games with a Purpose with Purpose (Dr Richard A. Bartle)*. Available at: https://www.youtube.com/watch?v=IeyZIs68AsA (Accessed: 23 October 2022).

Gartland, S., Flynn, P., Carneiro, M.A., Holloway, G., Fialho, J.d.S., Cullen, J., Hamilton, E., Harris, A. and Cullen, C. (2022) 'The state of web accessibility for people with cognitive disabilities: A rapid evidence assessment', *Behavioral Sciences*, 12(2), p. 26. Available at: https://doi.org/10.3390/bs12020026.

Gleason, C., Carrington, P., Cassidy, C., Morris, M.R., Kitani, K.M. and Bigham, J.P. (2019) 'It's almost like they're trying to hide it': How user-provided image descriptions have failed to make Twitter accessible', *The World Wide Web Conference*, pp. 549–559.

Gleason, C., Pavel, A., Liu, X., Carrington, P., Chilton, L.B. and Bigham, J.P. (2019) 'Making memes accessible', in *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 367–376.

Gleason, C., Pavel, A., McCamey, E., Low, C., Carrington, P., Kitani, K.M. and Bigham, J.P. (2020) 'Twitter A11y: A browser extension to make Twitter images accessible', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. Available at: https://doi.org/10.1145/3313831.3376728.

Green, N., Breimyer, P., Kumar, V. and Samatova, N. (2010) 'Packplay: Mining semantic data in collaborative games', in *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 227–234. Available at: https://aclanthology.org/W10-1837.pdf (Accessed: 2 February 2025).

Gregor, P., Newell, A.F. and Zajicek, M. (2002) 'Designing for dynamic diversity: interfaces for older people', in *Proceedings of the fifth international ACM conference on Assistive technologies*, pp. 151–156.

Griffith, M., Wentz, B. and Lazar, J. (2022) 'Quantifying the Cost of Web Accessibility Barriers for Blind Users', *Interacting with Computers*, 34(6), pp. 137–149.

Gubbi Mohanbabu, A. and Pavel, A. (2024) 'Context-aware image descriptions for web accessibility', in *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*, St. John's, NL, Canada. New York: ACM, pp. 1–17. Available at: https://doi.org/10.1145/3663548.3675658

Gubhka, M., Mohanbabu, A. and Pavel, A. (2024) 'Context-Aware Image Descriptions for Web Accessibility', in *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*, St. John's, NL, Canada: ACM, pp. 1–17. Available at: https://doi.org/10.1145/3663548.3675658.

Gudhka, M. (2021) ‘Good Alt Text, Bad Alt Text — Making Your Content Perceivable’, *WCAG*, 23 December. Available at: https://wcag.com/blog/good-alt-text-bad-alt-text-making-your-content-perceivable/ (Accessed: 12 May 2024).

Gwet, K.L. (2014) *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hackett, S., Parmanto, B. and Zeng, X. (2003) ‘Accessibility of Internet websites through time’, in *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 32–39.

Halpin, M., 2025. *Understanding the Barriers to Accessibility* [online]. Recite Me. Available at: https://reciteme.com/news/barriers-to-accessibility (Accessed 12 September 2025)

Hanley, M., Barocas, S., Levy, K., Azenkot, S. and Nissenbaum, H. (2021) ‘Computer Vision and Conflicting Values: Describing People with Automated Alt Text’, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES ’21: AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Event USA: ACM, pp. 543–554. Available at: https://doi.org/10.1145/3461702.3462620.

Harris, C. (2020) ‘ClueMeIn: Obtaining More Specific Image Labels Through a Game’, in *Workshop on Games and Natural Language Processing*, pp. 10–16.

Hecker, S., Bonney, R., Haklay, M., Hölker, F., Hofer, H., Goebel, C., Gold, M., Makuch, Z., Ponti, M., Richter, A. and Robinson, L. (2018) ‘Innovation in citizen science–perspectives on science-policy advances’, *Citizen Science: Theory and Practice*, 3(1), pp. 4–4.

Heylighen, A., Van der Linden, V. and Van Steenwinkel, I. (2017) ‘Ten questions concerning inclusive design of the built environment’,*Building and environment*, 114, pp. 507–517.

Hosking, I., Waller, S. and Clarkson, P.J. (2010) ‘It is normal to be different: Applying inclusive design in industry’, *Interacting with computers*, 22(6), pp. 496–501.

Howe, J. (2008) *Crowdsourcing: How the power of the crowd is driving the future of business*. New York: Random House.

Imrie, R. (2012) ‘Universalism, universal design and equitable access to the built environment’, *Disability and rehabilitation*, 34(10), pp. 873–882.

Iniesto, F., McAndrew, P., Minocha, S. and Coughlan, T. (2022) ‘A qualitative study to understand the perspectives of MOOC providers on accessibility’, *Australasian Journal of Educational Technology*, 38(1), pp. 87–101.

Ismailova, R. and Inal, Y. (2022) ‘Comparison of online accessibility evaluation tools: an analysis of tool effectiveness’, *IEEE Access*, 10, pp. 58233–58239.

ISO, I. and Guide, I.E.C. (2001) ‘71: Guidelines for standards developers to address the needs

of older persons and persons with disabilities'. ISO.

Iwarsson, S. and Ståhl, A. (2003) 'Accessibility, usability and universal design—positioning and definition of concepts describing person-environment relationships', *Disability and rehabilitation*, 25(2), pp. 57–66.

Johansson, S., Gulliksen, J. and Gustavsson, C. (2021) 'Disability digital divide: the use of the internet, smartphones, computers and tablets among people with disabilities in Sweden', *Universal Access in the Information Society*, 20(1), pp. 105–120. Available at: https://doi.org/10.1007/s10209-020-00714-x.

Jonsson, M., Johansson, S., Hussain, D., Gulliksen, J. and Gustavsson, C. (2023) 'Development and evaluation of eHealth services regarding accessibility: scoping literature review', *Journal of Medical Internet Research*, 25, e45118. Available at: https://doi.org/10.2196/45118

Kameda, T., Toyokawa, W. and Tindale, R.S. (2022) 'Information aggregation and collective intelligence beyond the wisdom of crowds', *Nature Reviews Psychology*, 1(6), pp. 345–357.

Kapur, R. and Kreiss, E. (2024) 'Reference-Based Metrics Are Biased Against Blind and Low-Vision Users' Image Description Preferences', in *Proceedings of the Third Workshop on NLP for Positive Impact*, pp. 308–314. Available at: https://aclanthology.org/2024.nlp4pi-1.26/ (Accessed: 21 May 2025).

Kaur, N. and Kumar, V. (2015a) 'Comparative analysis of automated web accessibility tools for developing and evaluating accessible websites', *FP-International Journal of Computer Science Research (IJCSR)*, 2(2), pp. 122–125. Available at: https://ijcsr.forexjournal.co.in/papers-pdf/29.pdf.

Kaur, N. and Kumar, V. (2015b) 'Framework for covering the limitations of web accessibility improvement tools', *IJCSR*, 2(1), pp. 27–31. Available at: https://ijcsr.forexjournal.co.in/papers-pdf/12.pdf.

Keates, S., Clarkson, P.J., Harrison, L.A. and Robinson, P. (2000) 'Towards a practical inclusive design approach', in *Proceedings on the 2000 conference on Universal Usability*, pp. 45–52.

Kelemen, M.L. and Rumens, N. (2008) 'An introduction to critical management research'. Available at: https://www.torrossa.com/gs/resourceProxy?an=4913318&publisher=FZ7200.

Kenigsberg, P.A., Aquino, J.P., Bérard, A., Brémond, F., Charras, K., Dening, T., Droës, R.M., Gzil, F., Hicks, B., Innes, A. and Nguyen, S.M. (2019) 'Assistive technologies to address capabilities of people with dementia: from research to practice', *Dementia*, 18(4), pp. 1568–1595.

Kercher, P. and EIDD, F.P. (2008) 'Design for All: changing the world by design', *4th Annual*

Issue, p. 27.

Kicikoglu, D., Bartle, R., Chamberlain, J. and Poesio, M. (2019) 'Wormingo: a'true gamification'approach to anaphoric annotation', in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pp. 1–7.

Kicikoglu, O.D., Bartle, R., Chamberlain, J., Paun, S. and Poesio, M. (2020) 'Aggregation driven progression system for GWAPs', in *Workshop on Games and Natural Language Processing*, pp. 79–84.

Kim, S., Yu, B., Li, Q. and Bolton, E.E. (2024) 'PubChem synonym filtering process using crowdsourcing', *Journal of Cheminformatics*, 16(1), p. 69. Available at: https://doi.org/10.1186/s13321-024-00868-3

Klaus, D., Haas, B. and Lamura, M. (2023) 'Dependency and Social Recognition of Online Platform Workers: Evidence From a Mixed-Methods Study', *Social Inclusion*, 11(4), pp. 251–261.

Krause, M. and Smeddinck, J. (2011) 'Human computation games: A survey', in *2011 19th European Signal Processing Conference*. IEEE, pp. 754–758.

Kreiss, E., Bennett, C., Hooshmand, S., Zelikman, E., Morris, M.R. and Potts, C. (2022) 'Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics', *arXiv preprint arXiv:2205.10646*. Available at: https://doi.org/10.48550/arXiv.2205.10646 (Accessed: 25 June 2025).

Kuppusamy, K.S. and Balaji, V. (2023) 'Evaluating web accessibility of educational institutions websites using a variable magnitude approach', *Universal Access in the Information Society*, 22(1), pp. 241–250. Available at: https://doi.org/10.1007/s10209-021-00812-4.

Kusmaryono, I., Wijayanti, D. and Maharani, H.R., 2022. 'Number of Response Options, Reliability, Validity, and Potential Bias in the Use of the Likert Scale: Education and Social Science Research – A Literature Review', *International Journal of Educational Methodology*, 8(4), pp.625–637.

Lafourcade, M. (2020) 'Game design evaluation of GWAPs for collecting word associations', in *Workshop on Games and Natural Language Processing*, pp. 26–33.

Lafourcade, M. and Le Brun, N. (2023) 'Apport du jeu pour la construction de connaissances: le projet jeuxdemots', *Technologie et innovation*, 8(4). Available at: https://openscience.fr/IMG/pdf/iste_techinn23v8n3_7.pdf (Accessed: 21 May 2025).

Lange, K. and Becerra, R. (2007) 'Teaching universal design in Colombia: The academic approach of two universities', in *Include 2007 conference proceedings*.

Launus-Gamble, C. (2021) *Writing Good Alt Text Image Descriptions*, KreativeInc Agency. Available at: https://kreativeincagency.co.uk/uncategorised/writing-good-alt-text-image-descriptions/ (Accessed: 2 February 2025).

Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D. and Cord, M. (2023) 'Obelics: An open web-scale filtered dataset of interleaved image-text documents', *Advances in Neural Information Processing Systems*, 36, pp. 71683–71702.

Laverde, M. (2021) *COVID 19, technology-based education and disability: the case of Colombia; emerging practices in inclusive digital learning for students with disabilities - UNESCO Digital Library*, Unesco. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000377773 (Accessed: 3 November 2023).

Law, E.L., Von Ahn, L., Dannenberg, R.B. and Crawford, M. (2007) 'TagATune: A Game for Music and Sound Annotation', in *ISMIR*, p. 2.

Lawson, A. (2011) 'Disability and employment in the Equality Act 2010: opportunities seized, lost and generated', *Industrial Law Journal*, 40(4), pp. 359–383.

Lee, H.-N. and Ashok, V. (2022) 'Impact of Out-of-Vocabulary Words on the Twitter Experience of Blind Users', in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–20.

Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.W. and Toutanova, K. (2023) 'Pix2struct: Screenshot parsing as pretraining for visual language understanding', in *International Conference on Machine Learning*. PMLR, pp. 18893–18912. Available at: https://proceedings.mlr.press/v202/lee23g.html (Accessed: 21 February 2025).

Lengua, C., Rubano, V. and Vitali, F. (2022) 'Aligning accessibility design to non-disabled people's perceptions', in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, pp. 1–6.

Leotta, M., Mori, F. and Ribaudo, M. (2023) 'Evaluating the effectiveness of automatic image captioning for web accessibility', *Universal Access in the Information Society*, 22(4), pp. 1293–1313.

Lichtman, M. (2013) 'Making meaning from your data', in *Qualitative Research in Education: A User's Guide*. 3rd edn. Thousand Oaks, CA: Sage, pp. 241–268. Available at: https://us.sagepub.com/sites/default/files/upm-binaries/45660_12.pdf (Accessed: 26 November 2023).

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick,

C.L. (2014) ‘Microsoft COCO: Common Objects in Context’, in D. Fleet et al. (eds.) *Computer Vision – ECCV 2014*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 740–755. Available at: https://doi.org/10.1007/978-3-319-10602-1_48.

Lorgat, M.G., Paredes, H. and Rocha, T. (2024) ‘A Gamification-Based Tool to Promote Accessible Design’, in A.K. Nagar et al. (eds.) *Intelligent Sustainable Systems*. Singapore: Springer Nature Singapore (Lecture Notes in Networks and Systems), pp. 373–390. Available at: https://doi.org/10.1007/978-981-99-8031-4_33.

Loseby, J. (2024) ‘The internet’s accessibility problem—and how to fix it’, *YouTube*. Available at: https://www.youtube.com/watch?v=QWPWgaDqbZI.

Luccioni, A.S. and Viviano, J.D. (2021) *What’s in the box? A preliminary analysis of undesirable content in the common crawl corpus*. arXiv preprint arXiv:2105.02732. Available at: https://doi.org/10.48550/arXiv.2105.02732

Lundgard, A. and Satyanarayan, A. (2022) ‘Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content’, *IEEE Transactions on Visualization and Computer Graphics*, 28(1), pp. 1073–1083. Available at: https://doi.org/10.1109/TVCG.2021.3114770.

Lyding, V., Nicolas, L. and König, A. (2022) ‘About the applicability of combining implicit crowdsourcing and language learning for the collection of NLP datasets’, in *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pp. 46–57. Available at: https://aclanthology.org/2022.nidcp-1.8/ (Accessed: 31 January 2024).

Mack, K., Cutrell, E., Lee, B. and Morris, M.R. (2021) ‘Designing Tools for High-Quality Alt Text Authoring’, in *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS ’21: The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, Virtual Event USA: ACM, pp. 1–14. Available at: https://doi.org/10.1145/3441852.3471207.

MacLeod, H., Bennett, C.L., Morris, M.R. and Cutrell, E. (2017) ‘Understanding blind people’s experiences with computer-generated captions of social media images’, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 5988–5999.

Madge, C., Yu, J., Chamberlain, J., Kruschwitz, U., Paun, S. and Poesio, M. (2019) ‘Progression in a language annotation game with a purpose’, in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pp. 77–85.

Madge, C., Brightmore, J., Kicikoglu, D., Althani, F., Bartle, R., Chamberlain, J., Kruschwitz,

U. and Poesio, M. (2022) ‘LingoTowns: A Virtual World For Natural Language Annotation and Language Learning’, in *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, pp. 57–62.

Madge, C.J. (2020) *Gamifying language resource acquisition*. PhD Thesis. Queen Mary University of London.

Malone, T.W., Laubacher, R. and Dellarocas, C. (2009) *Harnessing crowds: Mapping the genome of collective intelligence*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1381502 (Accessed: 5 January 2024).

Mangiatordi, A. and Lazzari, M. (2018) ‘Combined use of artificial intelligence and crowdsourcing to provide alternative content for images on websites’, in *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, pp. 1–6. Available at: https://ieeexplore.ieee.org/abstract/document/8319312/?casa_token=N6xuWFxL1EEAAAAA:2DTHmVErHoNcCClIdky4iEQHlqeFekWxaFFlQIJREBmfly6M-COUO52vO6mts-XOgTrVmJUK5a4.

Marcus-Quinn, A. (2022) ‘The EU Accessibility Act and Web Accessibility Directive and the implications for Digital Teaching and Learning Materials’, *Routledge Open Research*, 1(30), p. 30.

McCall, K. and Chagnon, B. (2022) ‘Rethinking Alt text to improve its effectiveness’, in *Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022, Proceedings, Part II*. Springer, pp. 26–33.

McCarthy, J.E. and Swierenga, S.J. (2010) ‘What we know about dyslexia and web accessibility: A research review’, *Universal Access in the Information Society*, 9(2), pp. 147–152.

McEwan, T. and Weerts, B. (2007) ‘ALT text and basic accessibility’ [Paper presentation]. Available at: https://doi.org/10.14236/ewic/HCI2007.64.

Miller, K. (2022) *Creating a More Accessible Internet: Context Matters | Stanford HAI*. Available at: https://hai.stanford.edu/news/creating-more-accessible-internet-context-matters (Accessed: 21 May 2025).

Miranda, D. and Araujo, J. (2022) ‘Studying industry practices of accessibility requirements in agile development’, in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pp. 1309–1317.

Mitchell, M. (2021) *Why AI is Harder Than We Think*. arXiv. Available at: https://doi.org/10.48550/arXiv.2104.12871.

Molina, J.L., Tubaro, P., Casilli, A. and Santos-Ortega, A. (2023) 'Research Ethics in the Age of Digital Platforms', *Science and Engineering Ethics*, 29(3), p. 17. Available at: https://doi.org/10.1007/s11948-023-00437-1.

Morash, V.S., Siu, Y.T., Miele, J.A., Hasty, L. and Landau, S. (2015) 'Guiding novice web workers in making image descriptions using templates', *ACM Transactions on Accessible Computing (TACCESS)*, 7(4), pp. 1–21.

Moreno, L., Martinez, P., Ruiz, B. and Iglesias, A. (2011) 'Toward an equal opportunity web: Applications, standards, and tools that increase accessibility', *Computer Magazine*, 44(5), pp. 18–26. Available at: https://doi.org/10.1109/MC.2010.370.

Moreno, L., Alarcón, R., Segura-Bedmar, I. and Martínez, P. (2019) 'Lexical simplification approach to support the accessibility guidelines', in Proceedings of the XX International Conference on Human Computer Interaction, pp. 1–4.

Morris, M.R., Zolyomi, A., Yao, C., Bahram, S., Bigham, J.P. and Kane, S.K. (2016) '"With most of it being pictures now, I rarely use it": Understanding Twitter's Evolving Accessibility to Blind Users', in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI'16: CHI Conference on Human Factors in Computing Systems*, San Jose California USA: ACM, pp. 5506–5516. Available at: https://doi.org/10.1145/2858036.2858116.

Morris, M.R., Johnson, J., Bennett, C.L. and Cutrell, E. (2018) 'Rich representations of visual content for screen reader users', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–11.

Mott, M., Cutrell, E., Franco, M.G., Holz, C., Ofek, E., Stoakley, R. and Morris, M.R. (2019) 'Accessible by Design: An Opportunity for Virtual Reality', in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Beijing, China: IEEE, pp. 451–454. Available at: https://doi.org/10.1109/ISMAR-Adjunct.2019.00122.

Muehlbradt, A. and Kane, S.K. (2022) 'What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning', *ACM Transactions on Accessible Computing (TACCESS)*, 15(1), pp. 1–32.

Nedelkina, M. (2022) 'Characteristics of an accessible web product and how to implement them: Recommendations for Brella Oy'.

Newell, A.F. and Gregor, P. (1999) 'Extra-Ordinary Human-Machine Interaction: What can be Learned from People with Disabilities?',*Cognition, Technology & Work*, 1(2), pp. 78–85.

Available at: https://doi.org/10.1007/s101110050034.

Nguyen, N.C., Pham, H.V., Wei, Z., Thawonmas, R., Paliyawan, P. and Harada, T. (2019) 'Using GWAP to retrieve informative description for Ukiyo-e images on live streaming platform', in *2019 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pp. 43–44. Available at: http://dx.doi.org/10.1109/ICCE-Asia46551.2019.8941600.

Nguyen, N.C., Thawonmas, R., Paliyawan, P. and Pham, H.V. (2020) 'JUSTIN: An audience participation game with a purpose for collecting descriptions for artwork images', in *2020 IEEE Conference on Games (CoG)*. IEEE, pp. 344–350.

Nielsen, J. (2000) *Why You Only Need to Test with 5 Users*. Nielsen Norman Group. Available at: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ (Accessed: 14 July 2025).

Noble, C. (2021) *Development of Web Accessibility Recommendations for CERN*. PhD Thesis. CERN. Available at: https://cds.cern.ch/record/2844654 (Accessed: 27 October 2023).

Nowell, L.S., Norris, J.M., White, D.E. and Moules, N.J. (2017) 'Thematic Analysis: Striving to Meet the Trustworthiness Criteria', *International Journal of Qualitative Methods*, 16(1), p. 160940691773384. Available at: https://doi.org/10.1177/1609406917733847.

O'Connell, T. and Goldberg, L. (2011) 'Universal design in media', in *Universal design handbook*. New York: McGraw-Hill.

Office of Disability Employment Policy (2022) 'Disability and the Digital Divide: Internet Subscriptions, Internet Use and Employment Outcomes'.

Open Inclusion (2022) *Attitudes to Digital Accessibility: Full Report November 2022*. AbilityNet. Available at: https://abilitynet.org.uk/attitudes2022 (Accessed: 1 February 2023).

Parasca, I.E., Rauter, A.L., Roper, J., Rusinov, A., Bouchard, G., Riedel, S. and Saito, P. (2016) 'Defining Words with Words: Beyond the Distributional Hypothesis', in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 122–126. Available at: https://doi.org/10.18653/v1/W16-2522.

Pascual, A., Ribera, M. and Granollers, T. (2014) 'Impact of web accessibility barriers on users with hearing impairment', in *Proceedings of the XV International Conference on Human Computer Interaction*, pp. 1–2.

Patton, M.Q. (1990) *Qualitative evaluation and research methods*. SAGE Publications, inc. Available at: https://psycnet.apa.org/record/1990-97369-000 (Accessed: 8 May 2024).

Persson, H., Åhman, H., Yngling, A.A. and 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, pp. 505–526.

Pe-Than, E.P.P., Goh, D.H.-L. and Lee, C.S. (2015) 'A typology of human computation games: an analysis and a review of current games', *Behaviour & Information Technology*, 34(8), pp. 809–824.

Petrie, H., Power, C.D., Swallow, D., Carlos, A.V., Gallagher, B., Magennis, M., Murphy, E., Sam, C. and Down, K. (2011) 'The value chain for web accessibility: Challenges and opportunities', in *Proceedings of Accessible Design in the Digital World 2011*.

Petrie, H. and Bevan, N. (2009) 'The evaluation of accessibility, usability, and user experience.', in *The universal access handbook*, 1, pp. 1–16.

Petrie, H., Höckner, K. and Rosenberger, W. (2022) 'Digital Accessibility: Readability and Understandability: Introduction to the Special Thematic Session', in *Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022, Proceedings, Part II*. Springer, pp. 3–5.

Petrie, H., Savva, A. and Power, C. (2015) 'Towards a unified definition of web accessibility', in *Proceedings of the 12th International Web for All Conference*, pp. 1–13.

Petrie, H.L., Weber, G. and Fisher, W. (2005) 'Personalization, interaction, and navigation in rich multimedia documents for print-disabled users', *IBM Systems Journal*, 44(3), pp. 629–635.

Petrosyan, A. (2023) *Internet penetration United States 2023*. Statista. Available at: https://www.statista.com/statistics/209117/us-internet-penetration/ (Accessed: 3 November 2023).

Power, C., Cairns, P. and Barlet, M. (2018) 'Inclusion in the third wave: Access to experience', in *New Directions in Third Wave Human-Computer Interaction: Volume 1-Technologies*. Cham: Springer, pp. 163–181.

Power, C., Freire, A., Petrie, H. and Swallow, D. (2012) 'Guidelines are only half of the story: accessibility problems encountered by blind users on the web', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.433–442.

Power, C. and Petrie, H. (2007) 'Accessibility in non-professional web authoring tools: a missed web 2.0 opportunity?', in *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, pp. 116–119.

Quinn, A.J. and Bederson, B.B. (2011) 'Human computation: a survey and taxonomy of a growing field', in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1403–1412.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S. and Finn, C. (2023) 'Direct

preference optimization: Your language model is secretly a reward model', *Advances in Neural Information Processing Systems*, 36, pp. 53728–53741. Available at: https://proceedings.neurips.cc/paper_files/paper/2023/file/9a1bf446d8c012949ff1af76ce62bc59-Paper-Conference.pdf (Accessed: 21 September 2025).

Ramos, R., Martins, B., Elliott, D. and Kementchedjhieva, Y. (2023) 'Smallcap: lightweight image captioning prompted with retrieval augmentation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2840–2849. Available at: http://openaccess.thecvf.com/content/CVPR2023/html/Ramos_SmallCap_Lightweight_Image_Captioning_Prompted_With_Retrieval_Augmentation_CVPR_2023_paper.html (Accessed: 21 February 2025).

Rao, K., Ok, M.W. and Bryant, B.R. (2014) 'A review of research on universal design educational models', *Remedial and Special Education*, 35(3), pp. 153–166.

Rashidi, M.N., Begum, R.A., Mokhtar, M. and Pereira, J.J. (2014) 'The conduct of structured interviews as research implementation method', *Journal of Advanced Research Design*, 1(1), pp. 28–34.

Ribot, J.C. and Peluso, N.L. (2003) 'A theory of access', *Rural sociology*, 68(2), pp. 153–181.

Rivero-Contreras, M., Engelhardt, P.E. and Saldaña, D. (2021) 'An experimental eye-tracking study of text adaptation for readers with dyslexia: effects of visual support and word frequency', *Annals of Dyslexia*, 71(1), pp. 170–187. Available at: https://doi.org/10.1007/s11881-021-00217-1.

Rodriguez, G. (2020) *Improving Web Accessibility Through Suggestions Using Serverless Architecture*. PhD Thesis. California State University, Northridge.

Rosenbaum, P. and Stewart, D. (2004) 'The World Health Organization International Classification of Functioning, Disability, and Health: a model to guide clinical thinking, practice and research in the field of cerebral palsy', in *Seminars in pediatric neurology*. Elsevier, pp. 5–10.

Rubya, S., Numainville, J. and Yarosh, S. (2021) 'Comparing Generic and Community-Situated Crowdsourcing for Data Validation in the Context of Recovery from Substance Use Disorders', in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21: CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, pp. 1–17. Available at: https://doi.org/10.1145/3411764.3445399.

Ruth-Janneck, D. (2011) 'Experienced barriers in web applications and their comparison to the WCAG guidelines', in *Information Quality in e-Health: 7th Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society,*

USAB 2011, Graz, Austria, November 25-26, 2011. Proceedings 7. Springer, pp. 283–300.

Sala, A., Arrue, M., Pérez, J.E. and Espín-Tello, S.M. (2020) 'Measuring complexity of e-government services for people with low vision', in *Proceedings of the 17th International Web for All Conference. W4A '20: 17th Web for All Conference*, Taipei Taiwan: ACM, pp. 1–5. Available at: https://doi.org/10.1145/3371300.3383350.

Salisbury, E., Kamar, E. and Morris, M. (2017) 'Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind', in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pp. 147–156.

Sanderson-Mann, J. and McCandless, F. (2005) 'Guidelines to the United Kingdom Disability Discrimination Act (DDA) 1995 and the Special Educational Needs and Disability Act (SENDA) 2001 with regard to nurse education and dyslexia', *Nurse Education Today*, 25(7), pp. 542–549.

Saunders, M., Lewis, P. and Thornhill, A. (2009) *Research methods for business students*. Pearson Education. Available at: https://books.google.com/books?hl=en&lr=&id=utxtfaCFiEC&oi=fnd&pg=PA2&dq=Research+Methods+for+Business+Students+(&ots=DyKTIoHe6N&sig=ayrw4glOimTy_NI3AD63pGoCDqk.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. and Schramowski, P. (2022) 'Laion-5b: An open large-scale dataset for training next generation image-text models', *arXiv preprint arXiv:2210.08402*.

Scott, M.J., Spyridonis, F. and Ghinea, G. (2015) 'Designing Accessible Games with the VERITAS Framework: Lessons Learned from Game Designers', in M. Antona and C. Stephanidis (eds.) *Universal Access in Human-Computer Interaction. Access to Learning, Health and Well-Being*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 547–554. Available at: https://doi.org/10.1007/978-3-319-20684-4_53.

Section508 (2023) *Section508.gov*. Available at: https://www.section508.gov/manage/laws-and-policies/ (Accessed: 11 December 2023).

Shakespeare, T. (2006) 'The social model of disability', *The disability studies reader*, 2, pp. 197–204.

Sharif, A., Chintalapati, S.S., Wobbrock, J.O. and Reinecke, K. (2021) 'Understanding screen-reader users' experiences with online data visualizations', in *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–16.

Sharma, P., Ding, N., Goodman, S. and Soricut, R. (2018) 'Conceptual captions: A cleaned,

hypernymed, image alt-text dataset for automatic image captioning', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565.

Shneiderman, B. (2022) *Human-centered AI*. Oxford University Press. Available at: https://books.google.com/books?hl=en&lr=&id=YS9VEAAAQBAJ&oi=fnd&pg=PP1&dq=Shneiderman,+B.+(2022)+Human+Centered+AI&ots=h2ueEMUM6a&sig=rK84mxHnTGzQiahEqf4h0LFZTMo (Accessed: 14 July 2025).

Silktide (2020) *When should you leave alt text blank? - Web accessibility FAQ - Silktide*. Available at: https://www.youtube.com/watch?v=VT4bDs1jMsA (Accessed: 2 February 2025).

Silktide (2023) *How to write good alt text - Making images accessible with alternative text*. Available at: https://www.youtube.com/watch?v=-6PxJxD08RI (Accessed: 26 February 2025).

Simons, R.N., Gurari, D. and Fleischmann, K.R. (2020) '"I Hope This Is Helpful": Understanding Crowdworkers' Challenges and Motivations for an Image Description Task', *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), pp. 1–26. Available at: https://doi.org/10.1145/3415176.

Singh, N., Wang, L.L. and Bragg, J. (2024) 'FigurA11y: AI Assistance for Writing Scientific Alt Text', in *Proceedings of the 29th International Conference on Intelligent User Interfaces. IUI '24: 29th International Conference on Intelligent User Interfaces*, Greenville SC USA: ACM, pp. 886–906. Available at: https://doi.org/10.1145/3640543.3645212.

Siu, K., Zook, A. and Riedl, M.O. (2017) 'A framework for exploring and evaluating mechanics in human computation games', in *Proceedings of the 12th International Conference on the Foundations of Digital Games*, pp. 1–4.

Spyridonis, F. and Daylamani-Zad, D. (2019) 'A serious game for raising designer awareness of web accessibility guidelines', in *IFIP Conference on Human-Computer Interaction*. Springer, pp. 3–12.

Spyridonis, F. and Daylamani-Zad, D. (2021) 'A serious game to improve engagement with web accessibility guidelines', *Behaviour & Information Technology*, 40(6), pp. 578–596. Available at: https://doi.org/10.1080/0144929X.2019.1711453.

Spyridonis, F., Daylamani-Zad, D. and Paraskevopoulos, I.Th. (2017) 'The gamification of accessibility design: A proposed framework', in *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*. 2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), Athens, Greece: IEEE, pp. 233–236. Available at: https://doi.org/10.1109/VS-GAMES.2017.8056606.

Srinivasan, K., Raman, K., Chen, J., Bendersky, M. and Najork, M. (2021) 'WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning', in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event Canada: ACM, pp. 2443–2449. Available at: https://doi.org/10.1145/3404835.3463257.

StackExchange (2017) *How exactly is sample size determined in inter-rater reliability study?* Available at: https://stats.stackexchange.com/questions/308953/how-exactly-is-sample-size-determined-in-inter-rater-reliability-study (Accessed: 11 December 2025).

Stangl, A., Verma, N., Fleischmann, K.R., Morris, M.R. and Gurari, D. (2021) 'Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision', in *Proceedings of the 23rd international ACM SIGACCESS conference on computers and accessibility*, pp. 1–15. Available at: https://dl.acm.org/doi/abs/10.1145/3441852.3471233?casa_token=lCO65xU1nPIAAAAA:OGZZOVlxPyuCbwTS3-4YiUkOyKY7pI-kZZMlf3pMGspBQlyTs1dlnm2zbZlumQ307XQV0CvTlw (Accessed: 21 May 2025).

Stangl, A., Morris, M.R. and Gurari, D. (2020) '"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', in *Proceedings of the 2020 chi conference on human factors in computing systems*, pp. 1–13.

Stangl, A.J., Kothari, E., Jain, S.D., Yeh, T., Grauman, K. and Gurari, D. (2018) 'BrowseWithMe: An Online Clothes Shopping Assistant for People with Visual Impairments', in *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS '18: The 20th International ACM SIGACCESS Conference on Computers and Accessibility*, Galway Ireland: ACM, pp. 107–118. Available at: https://doi.org/10.1145/3234695.3236337.

Steinfeld, E. and Danford, G.S. (1999) 'Theory as a basis for research on enabling environments', in *Enabling environments: Measuring the impact of environment on disability and rehabilitation*, pp. 11–33.

Steinmayr, B., Wieser, C., Kneißl, F. and Bry, F. (2011) 'Karido: A GWAP for telling artworks apart', in *2011 16th International Conference on Computer Games (CGAMES)*. IEEE, pp. 193–200.

Stratton, C., Kadakia, S., Balikuddembe, J.K., Peterson, M., Hajjioui, A., Cooper, R., Hong, B.-Y., Pandiyan, U., Munoz-Velasco, L.P., Joseph, J., Krassioukov, A., Tripathi, D.R. and Tuakli-Wosornu, Y.A. (2022) 'Access denied: The shortage of digitized fitness resources for

people with disabilities', *Disability and Rehabilitation*, 44(13), pp. 3301–3303. Available at: https://doi.org/10.1080/09638288.2020.1854873.

Suddaby, R. (2006) 'From the Editors: What Grounded Theory is Not', *Academy of Management Journal*, 49(4), pp. 633–642. Available at: https://doi.org/10.5465/amj.2006.22083020.

Surowiecki, J. (2005) *The wisdom of crowds*. Anchor. Available at: https://books.google.com/books?hl=el&lr=&id=hHUsHOHqVzEC&oi=fnd&pg=PR11&dq=wisdom+of+the+crowds+Surowiecki&ots=Zu7y0mVmcq&sig=BAf0-dYvKbPCizwmylFD3uLv7kw (Accessed: 6 January 2024).

Takagi, H., Kawanaka, S., Kobayashi, M., Sato, D. and Asakawa, C. (2009) 'Collaborative web accessibility improvement: challenges and possibilities', in *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pp. 195–202.

Taylor, P. (2022) *Global digital competitiveness country ranking 2022*. Statista. Available at: https://www.statista.com/statistics/1042743/worldwide-digital-competitiveness-rankings-by-country/ (Accessed: 3 November 2023).

Taylor, Z.W. and Bicak, I. (2019) 'Two-year institution and community college web accessibility: Updating the literature after the 2018 Section 508 amendment', *Community College Journal of Research and Practice*, 43(10–11), pp. 785–795.

Terry, G., Hayfield, N., Clarke, V. and Braun, V. (2017) 'Thematic analysis', in *The SAGE Handbook of Qualitative Research in Psychology*. 2nd edn. London: SAGE Publications, pp. 17–37.

Text descriptions working group and Cassidy, J. (2024) *Captions vs text descriptions*. BBC. Available at: https://www.bbc.co.uk/gel/features/the-difference-between-captions-and-text-descriptions (Accessed: 21 February 2025).

Thompson, J. (2022) 'A guide to abductive thematic analysis', *The Qualitative Report*, 27(5), pp. 1410–1421.

Tuite, K. (2014) 'GWAPs: Games with a problem.', in *FDG*.

Tunberg, M. (2022) 'Digital inclusion and the European Accessibility Act: both a necessity and an opportunity for TMT players'.

van der Smissen, D., Overbeek, A., van Dulmen, S., van Gemert-Pijnen, L., van der Heide, A., Rietjens, J.A. and Korfage, I.J. (2020) 'The feasibility and effectiveness of web-based advance care planning programs: scoping review', *Journal of Medical Internet Research*, 22(3), p. e15578. Available at: https://www.jmir.org/2020/3/e15578.

Vázquez, S.R. and Torres-del-Rey, J. (2019) 'Accessibility of multilingual information in

cascading crises', in *Translation in cascading crises*. Abingdon: Routledge, pp. 91–111.

Vickrey, D., Bronzan, A., Choi, W., Kumar, A., Turner-Maier, J., Wang, A. and Koller, D. (2008) 'Online word games for semantic data collection', in *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 533–542. Available at: https://aclanthology.org/D08-1056.pdf (Accessed: 2 February 2025).

Vigo, M., Brown, J. and Conway, V. (2013) 'Benchmarking web accessibility evaluation tools: Measuring the harm of sole reliance on automated tests', in *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pp. 1–10. Available at: https://doi.org/10.1145/2461121.2461124.

Vilares, J., Gómez-Rodríguez, C., Fernández-Núñez, L., Penas, D. and Viteri, J. (2020) 'Bringing Roguelikes to Visually-Impaired Players by Using NLP', in *Workshop on Games and Natural Language Processing*, pp. 59–67.

Vollenwyder, B., Buchmüller, E., Trachsel, C., Opwis, K. and Brühlmann, F. (2020) 'My train talks to me: participatory design of a mobile app for travellers with visual impairments', in *Computers Helping People with Special Needs: 17th International Conference, ICCHP 2020, Lecco, Italy, September 9–11, 2020, Proceedings, Part I 17*. Springer, pp. 10–18.

Vollenwyder, B., Petralito, S., Iten, G.H., Brühlmann, F., Opwis, K. and Mekler, E.D. (2023) 'How compliance with web accessibility standards shapes the experiences of users with and without disabilities', *International Journal of Human-Computer Studies*, 170, p. 102956.

Von Ahn, L. (2006) 'Games with a purpose', *Computer*, 39(6), pp. 92–94.

Von Ahn, L., Ginosar, S., Kedia, M., Liu, R. and Blum, M. (2006) 'Improving accessibility of the web with a computer game', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI06: CHI 2006 Conference on Human Factors in Computing Systems*, Montréal Québec Canada: ACM, pp. 79–82. Available at: https://doi.org/10.1145/1124772.1124785.

Von Ahn, L., Ginosar, S., Kedia, M. and Blum, M. (2007) 'Improving Image Search with PHETCH', in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, HI: IEEE, p. IV-1209-IV–1212. Available at: https://doi.org/10.1109/ICASSP.2007.367293.

Von Ahn, L. and Dabbish, L. (2004) 'Labeling images with a computer game', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319–326.

Von Ahn, L. and Dabbish, L. (2008) 'Designing games with a purpose', *Communications of the ACM*, 51(8), pp. 58–67.

W3C (1997) *World Wide Web Consortium (W3C) Launches International Web Accessibility Initiative*. W3C. Available at: https://www.w3.org/press-releases/1997/wai-launch/ (Accessed: 29 September 2023).

W3C (1999) *Techniques For Evaluation And Implementation Of Web Content Accessibility Guidelines*. Available at: https://www.w3.org/WAI/ER/IG/ert/ert-19991221.html (Accessed: 21 May 2025).

W3C (2016) *H67: Using null alt text and no title attribute on img elements for images that AT should ignore | Techniques for WCAG 2.0*. Available at: https://www.w3.org/TR/WCAG20-TECHS/H67.html (Accessed: 12 May 2024).

W3C (2023a) *H37: Using alt attributes on img elements | WAI | W3C*. Available at: https://www.w3.org/WAI/WCAG22/Techniques/html/H37 (Accessed: 5 May 2024).

W3C (2023b) *Understanding WCAG 2.0*. Available at: https://www.w3.org/TR/UNDERSTANDING-WCAG20/ (Accessed: 4 May 2024).

W3C (2024a) *An alt Decision Tree*. Web Accessibility Initiative (WAI). Available at: https://www.w3.org/WAI/tutorials/images/decision-tree/ (Accessed: 25 February 2025).

W3C (2024b) *Understanding Success Criterion 1.1.1: Non-text Content | WAI | W3C*. World Wide Web Consortium (W3C). Available at: https://www.w3.org/WAI/WCAG21/Understanding/non-text-content.html (Accessed: 24 February 2025).

W3C Web Accessibility Initiative (2023) *Home, Making the Web Accessible*. Available at: https://www.w3.org/WAI/ (Accessed: 11 December 2023).

Waller, S., Bradley, M., Hosking, I. and Clarkson, P.J. (2015) 'Making the case for inclusive design', *Applied ergonomics*, 46, pp. 297–303.

Wang, M.D. and Hau, K.T., 2025. 'In Likert Scale, Is Ticking Options Consecutively at Two Ends Equally Problematic as Ticking in the Middle?', *Journal of Official Statistics*, 41(1), pp.73–95.

WebAIM (2020) *WebAIM: The WebAIM Million - 2020 - An annual accessibility analysis of the top 1,000,000 home pages*. Available at: https://webaim.org/projects/million/2020 (Accessed: 29 September 2023).

WebAIM (2021) *WebAIM: Alternative Text*. Available at: https://webaim.org/techniques/alttext/ (Accessed: 19 April 2023).

WebAIM (2023) *WebAIM: The WebAIM Million - The 2023 report on the accessibility of the top 1,000,000 home pages*. Available at: https://webaim.org/projects/million/ (Accessed: 7 April 2023).

WebAIM (2024a) *WebAIM: Screen Reader User Survey #10 Results*. Available at: https://webaim.org/projects/screenreadersurvey10/ (Accessed: 1 May 2024).

WebAIM (2024b) *WebAIM: The WebAIM Million - The 2024 report on the accessibility of the top 1,000,000 home pages*. Available at: https://webaim.org/projects/million/ (Accessed: 30 April 2024).

WebAIM (2025) *WebAIM: The WebAIM Million - The 2025 report on the accessibility of the top 1,000,000 home pages*. Available at: https://webaim.org/projects/million/ (Accessed: 29 July 2025).

Weerasooriya, T.C., Luger, S., Poddar, S., KhudaBukhsh, A. and Homan, C. (2023) 'Subjective crowd disagreements for subjective data: Uncovering meaningful CrowdOpinion with population-level learning', in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July, pp. 950–966. Available at: https://doi.org/10.18653/v1/2023.acl-long.54

Williams, C., de Greef, L., Harris III, E., Findlater, L., Pavel, A. and Bennett, C. (2022) 'Toward supporting quality alt text in computing publications', in *Proceedings of the 19th International Web for All Conference*, pp. 1–12.

World Health Organization (2011) *World Report on Disability*. Available at: https://www.who.int/teams/noncommunicable-diseases/sensory-functions-disability-and-rehabilitation/world-report-on-disability (Accessed: 11 December 2023).

World Health Organization (2023) *Disability, Disability*. Available at: https://www.who.int/health-topics/disability (Accessed: 11 December 2023).

World Health Organization (2025) *Gender and health*. Available at: https://www.who.int/health-topics/gender (Accessed: 21 May 2025).

Wu, S., Wieland, J., Farivar, O. and Schiller, J. (2017) 'Automatic alt-text: Computer-generated image descriptions for blind users on a social network service', in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1180–1192.

Yeratziotis, A. and Zaphiris, P. (2018) 'A Heuristic Evaluation for Deaf Web User Experience (HE4DWUX)', *International Journal of Human–Computer Interaction*, 34(3), pp. 195–217. Available at: https://doi.org/10.1080/10447318.2017.1339940.

Yu, J., Paun, S., Camilleri, M., Garcia, P.C., Chamberlain, J., Kruschwitz, U. and Poesio, M. (2022) *Aggregating Crowdsourced and Automatic Judgments to Scale Up a Corpus of Anaphoric Reference for Fiction and Wikipedia Texts*. arXiv. Available at: https://doi.org/10.48550/arXiv.2210.05581.

Yuan, B., Sapre, M. and Folmer, E. (2010) ‘Seek-n-Tag: a game for labeling and classifying virtual world objects’, in *Proceedings of Graphics Interface 2010*. Citeseer, pp. 201–208.
Zhong, Y., Kobayashi, M., Matsubara, M. and Morishima, A. (2020) ‘Effects of crowd-in-the-loop alt text addition on the performance of visually impaired workers in online microtasks’, *The Transactions of Human Interface Society*, 22(3), pp. 251–262. Available at: https://doi.org/10.11184/his.22.3_251.
Zong, J., Lee, C., Lundgard, A., Jang, J., Hajas, D. and Satyanarayan, A. (2022) ‘Rich screen reader experiences for accessible data visualization’, *Computer Graphics Forum*, 41(3), pp. 15–27. Available at: https://doi.org/10.1111/cgf.14519.

Appendix A. Interview questions for visually impaired users (VIUs)

The core set of questions used in tandem with a web browsing task (see section 8.2) to guide discussions with VIUs are listed below. Not all were asked in every interview, but they rather acted as a guide and were adapted to the natural progression of each conversation.

A.1. Introduction questions

- Tell me about your experience in navigating the Web.
- How accessible do you feel web content is to you?
- What are the main challenges you face in navigating the Web?
- Do you do something to help deal with such challenges?

A.2. Screen readers and barriers

- Can you tell me about your experience using screen readers to navigate the Web?
- What are the main challenges you face in navigating the Web via screen readers?

A.3. Experience and expectations with alt text

- Can you tell me about your experience with alternative descriptions of web content like images?
- How satisfied are you with the quality of such descriptions on the Web?
- Do you feel that you could improve such descriptions if you could edit them?
- Do you have any specific expectations from such descriptions?
- All in all, what is the one thing that you feel is needed to improve the quality of such descriptions?

Appendix B. Interview questions for web content creators (WCCs)

The core set of questions used in tandem with a web browsing task (see section 8.2) to guide discussions with WCCs are listed below. Not all were asked in every interview, but they rather acted as a guide and were adapted to the natural progression of each conversation.

B.1. Introduction questions

- How long have you been involved in the creation of web content?
- Have you been involved in efforts to create accessible web content?
- Can you describe to me how do you go about creating accessible web content?
- How proficient would you say you are with creating accessible web content?
    - What do you think are the main benefits in focusing web design efforts towards accessibility?

      Or
    - What are the main reasons for not being involved much in the creation of accessible web content?

B.2. Barriers and WCAG

- Do you use any resources to increase your understanding in web accessibility?
- Are you familiar with web accessibility guidelines, such as WCAG?
    - To what extent do you aim to conform with such guidelines?
    - Do you think that conforming with such guidelines is sufficient to make web content accessible to all users?
- What do you think are the main challenges that people with disabilities or impairments face in navigating the Web?
- Do you do something to help surmount such challenges to make web content accessible to people with disabilities or impairments?

B.3. Experience and expectations with screen readers and alt text

What is your experience with screen readers:

- Have you, for example, created or evaluated web content specifically for being accessible to screen readers?
- What do you think are the main challenges that people who use screen readers face

when navigating the Web?

- How proficient would you say you are in writing good alt text descriptions?
- How effective do you believe alt text description that accompany web content are in describing such content?
- Do you have any key expectations from such descriptions to be of good quality?
- All in all, what is the one thing that you feel is needed to improve the quality of such descriptions?

Appendix C. GWAP backend: Database and entity relationship diagram

The structure of the database that was designed to expect the relevant data, that is, alt text and rating scores, submitted by the players of TagALTlong is shown in an Entity Relationship Diagram (ERD) (see Fig. C1 below).

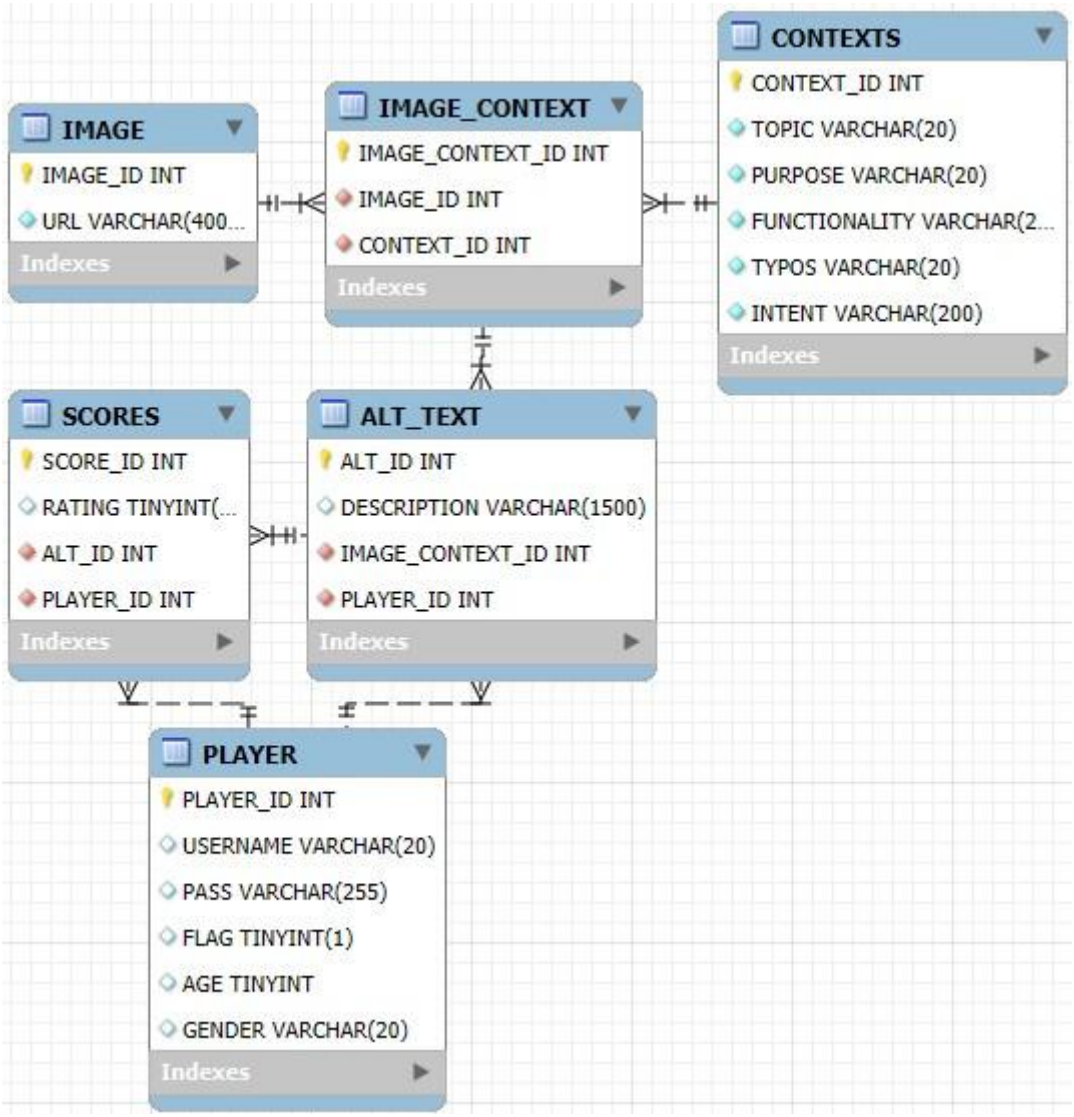


Fig. C1. Entity relationship diagram for the database of TagALTlong

As can be observed in the ERD, and for the purposes of data collection via TagALTlong, *one* (1) image can be used in *many* (N) contexts, while *one* (1) context can be applied to *many* (M) images, resulting in the creation of a table that represented this *many-to-many* (N-to-M) relationship between these entities, namely the 'IMAGE_CONTEXT' table in the above figure. *One* (1) combination of image and context can have *many* (N) alt text descriptions, as different

users can be presented with the same image-context combination when they play as authors (1-to-N relationship). This maps well to the **redundancy mechanism** from the typology presented in Table 9, allowing many players to author alt text for the same image-context combination and to also achieve **data refinement** through collective annotation for the same combination. Accordingly, *one* (1) player can author and rate *many* (N) alt text, creating two *one-to-many* relationships with the 'ALT_TEXT' and 'SCORES' tables, respectively; and, *one* (1) alt text can have *many* (N) scores, creating the final *one-to-many* (1-to-N) relationship between the latter tables. This relationship also relates to Pe-Than, Goh and Lee (2015)'s typology (Table 9), allowing players to collaboratively rate and refine alt text written by other players through **peer review and consensus**. Further, the fields for each of the tables were limited to essential information except for the 'FLAG' field in table 'PLAYER,' which was included to identify players that log in for the first time, so that the tutorial pops up to guide them before they are redirected to play as alt text authors.

Deferring to the database tables shown as entities in Fig. C1, new players complete the form in the register page (see Fig. 12.a), where they submit a username, a password, an age and a gender; and, these submissions will then populate the fields of table PLAYER, giving a value of '0' to the FLAG field until the player successfully logs in for the first time. For clarity, two such dataset entries were created by the researcher (see Fig. C2) to show how personal data submitted by players during registration, *e.g.,* account password, are not visible to owners of the database.

	PLAYER_ID	USERNAME	PASS	FLAG	AGE	GENDER
▶	1	Nick	$2y$10$iYCeDy5Xt3zRr9dttUm0remrEUER2OP…	1	23	Male
	2	Nikos	$2y$10$5JFoeIvpQm0ENg60AxawBuYGty1WQ…	1	27	Prefer not to say

Fig. C2. Example table player entries in the database

The entries in the above figure show how upon registration, player data in the database are only visible for non-personal information and that the password that players specify is hashed. This is achieved within the PHP backend code using the password_hash() function converting the plain text version of the password submitted by the user into an one-way hash before it is stored in the database. For clarity, the entries shown in Fig. C2 were not included in the dataset that was used to train the ML model and were only created as examples to show how player-submitted data are ultimately stored in the tables of the database.

Appendix D. GWAP frontend

Additionally, the game's frontend was implemented using Unity game engine and built as a WebGL game, allowing players to access it directly through a web browser. It was therefore hosted on itch.io [16], a popular platform for hosting indie video games, ensuring easy accessibility for players. Further, the game was implemented to work on average PCs, so that it is accessible in most places, and it was designed to be played using a mouse and a keyboard. In the backend, there is a total of 32 royalty-free sourced images and 1776 context descriptions, resulting in a pool of 20496 image-context combinations. Importantly, the game was carefully implemented to accommodate for the fact that not all combinations are used for all images, as some would not reasonably exist in an actual web context; for example, an image that is clearly not a graph will not have a context starting with 'This is a graph…'. This is very relevant in a GWAP context, as players have been shown to be discouraged to continue playing when it is clear that in-game content is non-applicable to real-world settings when otherwise implied (Vickrey *et al.*, 2008; Green *et al.*, 2010).

Accordingly, Fig. D1-D3 below demonstrate key functionalities at certain interaction points between the player and the backend. The figures are the corresponding code snippets of these functionalities, showing how image-context combinations are retrieved from the database and presented to players (Fig. D1), and how players successfully submit alt text (Fig. D2) and ratings (Fig. D3) to the database. For clarity, elements like UI handling, pop up management, error logging, and exact server endpoints are omitted from the code in the below figures.

[16] https://itch.io

```
public IEnumerator GetImageContext(RawImage randomImage, Text randomContext)
{
    string url = $"https://example.com/random_image_context.php?currentImageContextId={currentImageContextId}&playerId={currentPlayerId}";
    using (UnityWebRequest www = UnityWebRequest.Get(url))
    {
        yield return www.SendWebRequest();
        string response = www.downloadHandler.text.Trim();
        if (response == "MIN_AUTHORED")
        {
            ImageContextRandomizer randomizer = FindObjectOfType<ImageContextRandomizer>();
            if (randomizer != null && randomizer.minAuthoredPopUp != null)
            {
                randomizer.minAuthoredPopUp.SetActive(true);
            }
            yield break;
        }
        string[] lines = response.Split(new[] { '\n' }, System.StringSplitOptions.RemoveEmptyEntries);
        if (lines.Length >= 3)
        {
            currentImageContextId = int.Parse(lines[0]); // Store IMAGE_CONTEXT_ID
            string imageUrl = lines[1];
            string prompt = lines[2];
            using (UnityWebRequest imageRequest = UnityWebRequestTexture.GetTexture(imageUrl))
            {
                yield return imageRequest.SendWebRequest();
                Texture2D texture = DownloadHandlerTexture.GetContent(imageRequest);
                if (texture != null)
                {
                    randomImage.gameObject.SetActive(true);
                    randomImage.texture = texture;
                }
            }
            randomContext.text = prompt;
        }
    }
}
```

Fig. D1. GWAP code snippet for the GetImageContext coroutine: Retrieves and displays image-context combinations in the UI

The coroutine shown in the figure how the game retrieves a random image and its associated context prompt from the database (currentImageContextId) subject to user study requirements (see chapter 10), such as the need to author at least 10 alt text descriptions (MIN_AUTHORED) before rating alt text written by others. This is achieved with the retrieval of player information from the database (currentPlayerID). When this step is complete, the response is parsed and the image-context combination, including thr image URL (imageURL) and the context prompt (prompt), is extracted. Then, those are applied to the user interface (UI), with the image being assigned to the randomImage UI element, and the context prompt replacing the text found in the randomContext UI element.

```
public IEnumerator NewAltText(string altText, int imageContextId, System.Action onAltSubmitted)
{
    WWWForm form = new WWWForm();
    form.AddField("submittedAlt", altText);
    form.AddField("imageContextId", imageContextId);
    form.AddField("playerId", currentPlayerId);

    using (UnityWebRequest www = UnityWebRequest.Post("https://example.com/new_alt_text.php", form))
    {
        yield return www.SendWebRequest();
        string response = www.downloadHandler.text;
        if (response.StartsWith("success;"))
        {
            currentUsername = response.Split(';')[1];
            onAltSubmitted?.Invoke();
        }
    }
}
```

Fig. D2. GWAP code snippet for the NewAltText coroutine: Handles the submission of newly authored alt text by players

The coroutine in Fig. D2 shows how the submission of alt text by players is handled. Again, player (currentPlayerId) and image-context (imageContextId) information are retrieved from the database. These are added as fields to a WWWForm (imageContextID, playerId), so that the server can recognise the submission once the player submits an alt text (submittedAlt) for this image-context combination. The request to submit the alt text to the database is then sent to the server (UnityWebRequest.Post) and once the request is successful, the player's username (currentUsername) is extracted for the game to display it locally through a thank-you-pop-up, and the UI is updated loading a new image-context combination (onAltSubmitted?.Invoke()).

```
public IEnumerator NewScore(int score, System.Action onSuccess)
{
    WWWForm form = new WWWForm();
    form.AddField("rating", score); // Add the score to be submitted
    form.AddField("altId", currentAltId); // Add the current ALT_ID
    form.AddField("playerId", currentPlayerId); // Add the current PLAYER_ID

    using (UnityWebRequest www = UnityWebRequest.Post("https://example.com/new_score.php", form))
    {
        yield return www.SendWebRequest();
            string response = www.downloadHandler.text; // Server response
            if (response.StartsWith("success;"))
            {
                currentUsername = response.Split(';')[1]; // Extract the username
                onSuccess?.Invoke();
            }
    }
}
```

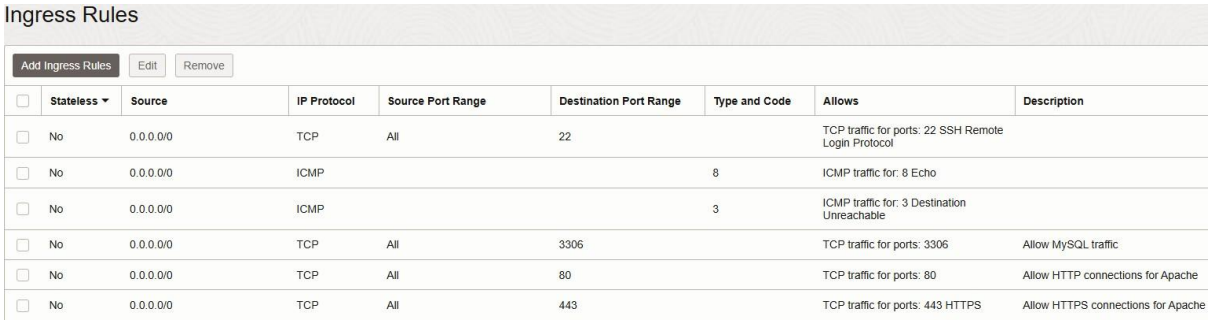
Fig. D3. GWAP code snippet for the NewScore coroutine: Handles the submission of new ratings by players

The coroutine shown in Fig. D3 shows how the submission of new rating scores by players are handled and stored in the database. First, a WWWform is used to add fields for the rating score (rating), as well as identifiers for the alt text the rating corresponds to (altId) and the player that

submitted the rating for this alt text. The procedure is then similar to the submission of new alt text (Fig. D2); indicatively, the form is sent via a POST request to the PHP backend script (Fig. E4) that will store the newly submitted rating score in the database. Finally, the username of the player (currentUsername) is again extracted and displayed in-game to thank the player, and a callback (onSuccess?.Invoke()) to update the game's UI.

Appendix E. GWAP PHP backend

First, ingress rules needed to be set for ports 80, 443 and 3306 on the OCI dashboard allowing Apache connections to HTTP and HTTPS, and TCP traffic, respectively (see Fig. E1 below).



Ingress Rules

Add Ingress Rules | Edit | Remove

Stateless	Source	IP Protocol	Source Port Range	Destination Port Range	Type and Code	Allows	Description
No	0.0.0.0/0	TCP	All	22		TCP traffic for ports: 22 SSH Remote Login Protocol	
No	0.0.0.0/0	ICMP			8	ICMP traffic for: 8 Echo	
No	0.0.0.0/0	ICMP			3	ICMP traffic for: 3 Destination Unreachable	
No	0.0.0.0/0	TCP	All	3306		TCP traffic for ports: 3306	Allow MySQL traffic
No	0.0.0.0/0	TCP	All	80		TCP traffic for ports: 80	Allow HTTP connections for Apache
No	0.0.0.0/0	TCP	All	443		TCP traffic for ports: 443 HTTPS	Allow HTTPS connections for Apache

Fig. E1. OCI dashboard ingress rules

The above figure shows how the ingress rules were set up in the OCI dashboard allowing for the aforementioned traffic and connections for the relevant ports. Then, traffic also needed to be allowed through the firewall on the Ubuntu VM itself; and, a MySQL bind address, as well as MySQL account with remote access to said address needed to be set up followed by a test of the PHP application on a local machine using *XAMPP*, ensuring that the application worked properly in a local environment before moving on to a production environment. Then, PHP and Apache needed to be installed in the same VM, and the latter was configured for PHP and to connect to MySQL before configuring the firewall and setting permissions accordingly.

As previously explained, the PHP backend serves as a communication layer between the GWAP and the database, given that the latter is hosted in a remote server provided by OCI. This is also the reason that the PHP files needed to be installed on the same Ubuntu VM as the database alongside the Apache web server. The frontend C# files are coded in a way allowing both the submission of alt text and rating scores by players in the database, as well as to request for data from the database, e.g., image-context combinations, player usernames, etc., through the PHP backend files. The latter files, then, handle the storing of player-submitted data in the database and the fetching of already submitted data to send them back to the frontend C# files. More colloquially, and upon successful configuration on the Ubuntu VM, the database accepts data storage operations from the backend and provides data retrieval operations to the backend, which the latter uses to communicate with the frontend (see Fig. 10). For completeness, the PHP code snippets relating to the previously presented functionalities — retrieve and display image-context combinations (Fig. D1) and submit new alt text (Fig. D2) — are presented below.

```php
<?php
// Step 1: Fetch a random IMAGE_CONTEXT_ID
$randomContextIdSql = "SELECT IMAGE_CONTEXT_ID FROM IMAGE_CONTEXT WHERE IMAGE_CONTEXT_ID != ? ORDER BY RAND() LIMIT 1;";
$stmt = $conn->prepare($randomContextIdSql);
$stmt->bind_param("i", $currentImageContextId); // Bind the current IMAGE_CONTEXT_ID
$stmt->execute();
$randomContextResult = $stmt->get_result();

if ($randomContextResult && $randomContextResult->num_rows > 0) {
    $randomContextRow = $randomContextResult->fetch_assoc();
    $imageContextId = $randomContextRow['IMAGE_CONTEXT_ID'];

    // Step 2: Fetch the image and context details
    $sql = "SELECT ic.IMAGE_CONTEXT_ID, i.URL,
            CONCAT('This ', c.TYPOS, ' ', c.FUNCTIONALITY, ' on ', c.TOPIC, ' webpage with the goal of ', c.PURPOSE, '. The intention of the image is to ', c.INTENT, '.') AS prompt
        FROM IMAGE_CONTEXT ic JOIN IMAGE i ON ic.IMAGE_ID = i.IMAGE_ID JOIN CONTEXTS c ON ic.CONTEXT_ID = c.CONTEXT_ID WHERE ic.IMAGE_CONTEXT_ID = ?";
    $stmt = $conn->prepare($sql);
    $stmt->bind_param("i", $imageContextId);
    $stmt->execute();
    $result = $stmt->get_result();

    if ($result && $result->num_rows > 0) {
        $row = $result->fetch_assoc();
        echo $row['IMAGE_CONTEXT_ID'] . "\n";
        echo $row['URL'] . "\n";
        echo $row['prompt'] . "\n";
    } else {
        echo "No contexts found in the database.";
    }
} else {
    echo "No IMAGE_CONTEXT_ID found.";
}
?>
```

Fig. E2. PHP backend code snippet: Retrieves image-context data from the database

This PHP script is called by the coroutine shown in Fig. D1 and aims to retrieve image-context information from the database and return them to the frontend to update the UI of the GWAP accordingly. First, it uses a SELECT query to retrieve a random image-context combination (IMAGE_CONTEXT_ID) to ensure that a new combination is displayed to the player every time. This is followed by a second, more complex query that retrieves relevant unique image- and context-specific data, relating to this IMAGE_CONTEXT_ID. Finally, the retrieved data are formatted appropriately and returned to the frontend (echo $row['IMAGE_CONTEXT_ID'] . "\n";, echo $row['URL'] . "\n";, and echo $row['prompt'] . "\n";). For clarity, certain parts of the script that did not relate to its key functionality were removed (e.g., the MIN_AUTHORED check, error handling, database connection details).

```php
<?php
// Step 1: Check if the player already submitted an alt text for this image context
$checkExistingAltSql = "SELECT ALT_ID FROM ALT_TEXT WHERE IMAGE_CONTEXT_ID = ? AND PLAYER_ID = ?";
$stmt = $conn->prepare($checkExistingAltSql);
$stmt->bind_param("ii", $imageContextId, $playerId);
$stmt->execute();
$checkResult = $stmt->get_result();

if ($checkResult->num_rows > 0) {
    echo "error;You have already submitted alt text for this image.";
    exit();
}

// Step 2: Insert the new alt text
$sql = "INSERT INTO ALT_TEXT (DESCRIPTION, IMAGE_CONTEXT_ID, PLAYER_ID) VALUES (?, ?, ?)";
$stmt = $conn->prepare($sql);
$stmt->bind_param("sii", $newAlt, $imageContextId, $playerId); // Bind the parameters

if ($stmt->execute()) {
    // Fetch the username after inserting the alt text
    $usernameSql = "SELECT USERNAME FROM PLAYER WHERE PLAYER_ID = ?";
    $usernameStmt = $conn->prepare($usernameSql);
    $usernameStmt->bind_param("i", $playerId);
    $usernameStmt->execute();
    $usernameResult = $usernameStmt->get_result();

    if ($usernameResult && $usernameResult->num_rows > 0) {
        $usernameRow = $usernameResult->fetch_assoc();
        $username = $usernameRow['USERNAME'];
        echo "success;{$username}"; // Return the success response with username
    } else {
        echo "error;Username not found"; // If the username isn't found, return error
    }
}
else { echo "error;{$stmt->error}"; }
?>
```

Fig. E3. PHP backend code snippet: Insert new alt text data to the database

This PHP script is called by the coroutine shown in Fig. D2 and aims to validate to store the alt text data submitted by players through the GWAP UI in the database. First, it uses a SELECT query to check against duplicate entries, and if an alt text has yet to be submitted by this player for this image-context combination, a new INSERT query is executed to store the information in the database. Finally, the username of the player is retrieved, so that a confirmation message can be displayed in the GWAP UI. For clarity, certain parts of the script that did not relate to its key functionality were removed (e.g., database connections and variable declarations).

```php
<?php
$rating = $_POST['rating'];
$altId = $_POST['altId'];
$playerId = $_POST['playerId'];

$checkSql = "SELECT RATING FROM SCORES WHERE ALT_ID = ? AND PLAYER_ID = ?";
$checkStmt = $conn->prepare($checkSql);
$checkStmt->bind_param("ii", $altId, $playerId);
$checkStmt->execute();
$checkResult = $checkStmt->get_result();

    $insertSql = "INSERT INTO SCORES (RATING, ALT_ID, PLAYER_ID) VALUES (?, ?, ?)";
    $insertStmt = $conn->prepare($insertSql);
    $insertStmt->bind_param("iii", $rating, $altId, $playerId);

    if ($insertStmt->execute()) {
        // Fetch the username
        $usernameSql = "SELECT USERNAME FROM PLAYER WHERE PLAYER_ID = ?";
        $usernameStmt = $conn->prepare($usernameSql);
        $usernameStmt->bind_param("i", $playerId);
        $usernameStmt->execute();
        $usernameResult = $usernameStmt->get_result();

        if ($usernameResult && $usernameResult->num_rows > 0) {
            $usernameRow = $usernameResult->fetch_assoc();
            $username = $usernameRow['USERNAME'];
            echo "success;{$username}";
        } else {
            echo "success;Unknown";
        }

        $usernameStmt->close();
}
$checkStmt->close();
$conn->close();
?>
```

Fig. E4. PHP backend code snippet: Insert new rating score to the database

This PHP script is called by the NewScore coroutine (Fig. D3) and aims to validate and store the rating scores submitted by players for a specific alt text through the game in the database. First, the POST variables (rating, altId, playerId) are retrieved and a SELECT query is used to check if this player has already submitted a rating score for this alt text. If there no duplicates, an INSERT query is used to store the new rating score in table SCORES and the username of the player is fetched and returned to the frontend (echo “success;{$username}”). For clarity, certain parts of the script that did not relate to its key functionality were removed (e.g., database connection details and security headers).

Appendix F. Online survey format sample

Rate how suitable each alt text description is for images used in specific contexts—We're counting on you!

Guidelines for suitable alt text

- Alt text doesn't complement an image, it's an alternative to an image.
- We don't read alt text. We listen to alt text via software known as screen readers.
- If an image is just there for the visual appeal, it should be marked as 'Eye Candy'.
- If it's not an 'Eye Candy', start with: 'How would you describe it to somebody over the phone?'
- Alt text shouldn't exceed 150 characters.
- Alt text shouldn't say 'An image of...', 'A picture of...', etc.
- Alt text should mention the functionality of the image if any (e.g. Link, Button, Logo, Icon).

The above are included at the top of every section of the survey and in each section, participants complete the below Likert scale 1-5 for each image, context prompt, and alt text tuple:

Context: This photograph is found on a social media webpage with the goal of advertising. The intention of the image is to guide (e.g., instructions or steps).
Alt text: **coffee beans and a coffee pot**

How suitable is the alt text for this image in this context?
1-Not suitable at all, 2-Not very suitable, 3-It's okay, 4-Suitable, 5-Very suitable

Appendix G. Latest ethics approval letter

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University of London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

6 February 2025

LETTER OF APPROVAL

APPROVAL HAS BEEN GRANTED FOR THIS STUDY TO BE CARRIED OUT BETWEEN 06/02/2025 AND 30/09/2025

Applicant (s): Mr Nikolaos Droutsas

Project Title: Purposeful Game Solutions for the Enjoyable Construction of Linguistic Resources to Improve Web Accessibility

Reference: 41665-A-Feb/2025- 53701-1

Dear Mr Nikolaos Droutsas

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has agreed that there is no objection on ethical grounds to the proposed study. Approval is given on the understanding that the conditions of approval set out below are followed:

- **The agreed protocol must be followed. Any changes to the protocol will require prior approval from the Committee by way of an application for an amendment.**
- **Please be advised that recruitment via social media is limited to you creating / sharing a post on your own channels, which participants can respond to or share if they are interested. It does not extend to direct messaging of large numbers of contacts, which can be considered spam.**

Please note that:

- Research Participant Information Sheets and (where relevant) flyers, posters, and consent forms should include a clear statement that research ethics approval has been obtained from the relevant Research Ethics Committee.
- The Research Participant Information Sheets should include a clear statement that queries should be directed, in the first instance, to the Supervisor (where relevant), or the researcher. Complaints, on the other hand, should be directed, in the first instance, to the Chair of the relevant Research Ethics Committee.
- Approval to proceed with the study is granted subject to any conditions that may appear above.
- The Research Ethics Committee reserves the right to sample and review documentation, including raw data, relevant to the study.
- If your project has been approved to run for a duration longer than 12 months, you will be required to submit an annual progress report to the Research Ethics Committee. You will be contacted about submission of this report before it becomes due.
- You may not undertake any research activity if you are not a registered student of Brunel University or if you cease to become registered, including abeyance or temporary withdrawal. As a deregistered student you would not be insured to undertake research activity. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

Appendix H. Latest participant information sheet (PIS)

PARTICIPANT INFORMATION SHEET

Study title

Purposeful Game Solutions for the Enjoyable Construction of Linguistic Resources to Improve Web Accessibility

Invitation Paragraph

You are being asked to take part in this research study. Before you decide, it is important for you to understand why this research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me/us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

What is the purpose of the study?

This study is being undertaken in the course of the doctoral, PhD, degree in Computer Science at Brunel University London, and its aim and background are as follows:

Study Aim: The aim of this research is to explore the potential of games to construct a thesaurus of alt text descriptions, that is, alternative descriptions of non-textual content across digital media, which could be used in a range of media to improve digital accessibility.

Study Background: Despite recent acknowledgements in research on digital media accessibility about the importance of inclusive web design, the lack and/or poor quality of alt text descriptions for images, videos, infographics, extended reality (XR), among others, has been noted by researchers. It has been argued however, that such descriptions need not only be present and accurate, but also clear and of sufficient volume to be useful, namely suitable alt text. Researchers however, highlight that suitable alt text authorship is an, admittedly, tedious and laborious task, and yet no gaming solution to this day, has, to the best of my knowledge, been developed and/or used for suitable alt text authorship; despite games' acclaimed potential in training and engaging players with typically laborious tasks such as the previously mentioned suitable alt text authorship.

Why have I been invited to participate?

General Population (20 participants in total)
Inclusion Criteria — **1)** At least 18 years old; **2)** Digital literacy, ability to, comfortably, use digital technology and/or devices.

Do I have to take part?

As participation is entirely voluntary, it is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and you may be asked to sign a consent form. If you decide to take part you are still free to withdraw at any time up until **30/06/2025** and without having to give a reason.

What will happen to me if I take part?

You will be asked to take part in interviews and/or playtest a digital game for approximately forty-five minutes. You might also be asked to complete a questionnaire, as part of a survey, which will not last more than twenty minutes. The interviews, playtests and/or questionnaires that you will be asked to take part in, as part of this study, will be taking place online.

Are there any lifestyle restrictions?

Please rest assured that there are no life style restrictions required for this study.

What are the possible disadvantages and risks of taking part?

For this study, the following risks and measures to mitigate them are in place:

Risks of distress, anxiety and psychological harm

Although such risks are neither very likely to take place, nor expected to cause serious discomfort, as part of the playtest iterations, the semi-structured interviews and/or the survey, I wish to let you know that if you, as the participant, ever feel the need to pause, stop or withdraw from this research at any time, due to being exposed to any of the afore-mentioned discomforting conditions, you are always free to do so and without having to give a reason. However, you are kindly requested to declare any disability and/or medical condition that you have been, knowingly, suffering from, so that the PhD researcher might be in a position to ensure your health and safety to the best of their effort and ability. If you have visual impairments, please rest assured that you will not be asked to read or write any information as part of this study; conversely, you will only be asked questions verbally.

Since the study will be taking place online, no further disadvantages and/or risks are anticipated for this research. Please rest assured that all necessary precautions as per Brunel University's guidelines will be followed.

What are the possible benefits of taking part?

As the study aims to explore gaming solutions to the generation of more suitable alt text descriptions, in the future, you may directly benefit from such descriptions as their quality and presence, researchers suggest, will reduce the interaction burden experienced by users with visual impairments while browsing the web and/or other forms of digital media. The general population is also expected to benefit directly from the outcome of this research, as there is evidence within accessibility studies to suggest that if digital media become more accessible to users with various forms of impairments, users without such impairments are highly likely to experience increased usability and/or user experience while interacting with digital media. Finally, direct benefits are also envisioned for web professionals, as the study will provide an enjoyable, as in gameful, solution for such professionals to engage with the generation of suitable alternative text descriptions, and recent research suggests that the resources available to web professionals to engage with creation of digital accessible content are few and unengaging.

What if something goes wrong?

Although we do not expect that something will be going wrong during your participation in this online study; however, if you, at any point, feel unhappy by taking part in this study, the person to be contacted if you wish to complain about the experience is **Professor Simon Taylor -** simon.taylor@brunel.ac.uk (the Chair of the College of Engineering, Design and Physical Sciences Research Ethics).

Will my taking part in this study be kept confidential?

All information which is collected about you during the course of the research will be kept strictly confidential for at least 10 years, following the completion of the study, and any information about you which leaves the University will have all your identifying information removed. With your permission, anonymised data will be stored and may be used in future research – you can indicate whether or not you give permission for this by way of the Consent Form. If during the course of the research evidence

of harm or misconduct come to light, then it may be necessary to break confidentiality. We will tell you at the time if we think we need to do this, and let you know what will happen next.

Will I be recorded, and how will the recording be used?

Interviews and playtesting sessions will be video recorded, and in such a case, they will be destroyed once transcribed. The recordings will not be capturing your face and/or features, as they are only concerned with your interaction with the game, if playtesting, or your verbal answers during the interviews and/or surveys, so that the data can then be transcribed.

What will happen to the results of the research study?

The results of this research will be written up as part of a PhD thesis, and parts of it may be published in Brunel University's Library, peer-reviewed scientific journals, conference presentations, internal reports, and/or written feedback to participants; however, participants will not be able to be identified in any report and/or publication unless they specifically request to do so. Copies of publications, relating to the results obtained within this research, will, upon their publication, be made available online and in Brunel University's Library. Research participants may be informed of the results of this study, once established, upon request.

Who is organising and funding the research?

This research is being organised by Mr Nikolaos Droutsas (Doctoral Researcher) in conjunction with Brunel University London.

What are the indemnity arrangements?

Brunel University London provides appropriate insurance cover for research which has received ethical approval.

Who has reviewed the study?

College of Engineering, Design and Physical Sciences Research Ethics Committee

Chair – Professor Simon Taylor (Simon.Taylor@brunel.ac.uk)

Research Integrity

Brunel University London is committed to compliance with the Universities UK Research Integrity Concordat. You are entitled to expect the highest level of integrity from the researchers during the course of this research.

Contact for further information and complaints

Researcher name and details: **Nikolaos Droutsas** - nick.droutsas@brunel.ac.uk

Supervisor name and details: **Fotios Spyridonis** - fotios.spyridonis@brunel.ac.uk

For complaints, Chair of the Research Ethics Committee: **Simon Taylor -** simon.taylor@brunel.ac.uk

---

You, the participant, will be given a copy of this *Participant Information Sheet* and a *signed consent form* to keep.

Thank you for reading through this document and for considering to support my research.

Appendix I. Latest consent form

CONSENT FORM

Purposeful Game Solutions for the Enjoyable Construction of Linguistic Resources to Improve Web Accessibility

Nikolaos Droutsas

APPROVAL HAS BEEN GRANTED FOR THIS STUDY TO BE CARRIED OUT BETWEEN 01/11/2023 AND 30/09/2025

The participant (or their legal representative) should complete the whole of this sheet.		
	YES	NO
Have you read the Participant Information Sheet?	☐	☐
Have you had an opportunity to ask questions and discuss this study? (via email/phone for electronic surveys)	☐	☐
Have you received satisfactory answers to all your questions? (via email/phone for electronic surveys)	☐	☐
Who have you spoken to about the study?		
Do you understand that you will not be referred to by name in any report concerning this study?	☐	☐
Do you understand that:		
• You are free to withdraw from this study at any time	☐	☐
• You don't have to give any reason for withdrawing	☐	☐
• Choosing not to participate or withdrawing will not affect your rights	☐	☐
• You can withdraw your data any time up to 30/06/2025	☐	☐
I agree to the use of non-attributable quotes when the study is written up or published	☐	☐
The procedures regarding confidentiality have been explained to me	☐	☐
I agree that my anonymised data can be stored and shared with other researchers for use in future projects.	☐	☐
I agree to take part in this study.	☐	☐

Signature of research participant:	
Print name:	Date: