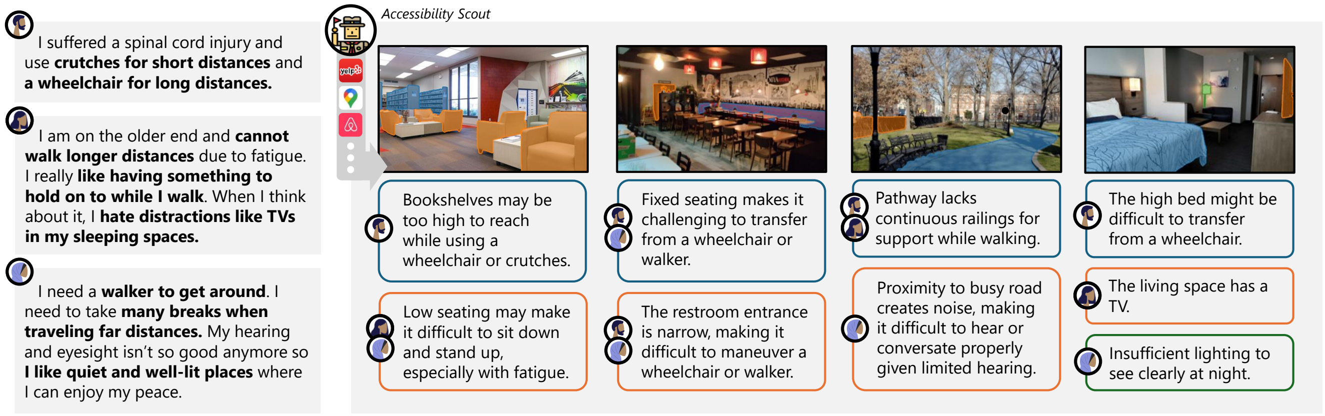
Accessibility Scout: Personalized Accessibility Scans of Built Environments

William Huang
University of California, Los Angeles
Los Angeles, CA, USA
william.huang@ucla.edu

Xia Su
University of Washington
Seattle, WA, USA
xiasu@cs.washington.edu

Jon E. Froehlich
University of Washington
Seattle, WA, USA
jonf@cs.uw.edu

Yang Zhang
University of California, Los Angeles
Los Angeles, CA, USA
yangzhang@ucla.edu



I suffered a spinal cord injury and use **crutches for short distances** and **a wheelchair for long distances.**

I am on the older end and **cannot walk longer distances** due to fatigue. I really **like having something to hold on to while I walk**. When I think about it, I **hate distractions like TVs in my sleeping spaces.**

I need a **walker to get around**. I need to take **many breaks when traveling far distances.** My hearing and eyesight isn't so good anymore so **I like quiet and well-lit places** where I can enjoy my peace.

Bookshelves may be too high to reach while using a wheelchair or crutches.

Low seating may make it difficult to sit down and stand up, especially with fatigue.

Fixed seating makes it challenging to transfer from a wheelchair or walker.

The restroom entrance is narrow, making it difficult to maneuver a wheelchair or walker.

Pathway lacks continuous railings for support while walking.

Proximity to busy road creates noise, making it difficult to hear or conversate properly given limited hearing.

The high bed might be difficult to transfer from a wheelchair.

The living space has a TV.

Insufficient lighting to see clearly at night.

Figure 1: Accessibility Scout is an LLM-based personalized accessibility scanning system for semi-automatically *modeling* a person's accessibility preferences to *identify* and *visualize* accessibility concerns in images of built environments. (Left) Accessibility Scout converts plain text descriptions of accessibility and collaborative human-AI accessibility annotations into LLM-interpretable user models. (Right) User models are used to generate personalized accessibility scans for each individual. Images can be sourced from anywhere, including Yelp, Google Maps, AirBnB, Facebook, Booking.com, and more.

Abstract

Assessing the accessibility of unfamiliar built environments is critical for people with disabilities. However, manual assessments, performed by users or their personal health professionals, are laborious and unscalable, while automatic machine learning methods often neglect an individual user's unique needs. Recent advances in Large Language Models (LLMs) enable novel approaches to this problem, balancing personalization with scalability to enable more adaptive and context-aware assessments of accessibility. We present Accessibility Scout, an LLM-based accessibility scanning system that identifies accessibility concerns from photos of built environments. With use, Accessibility Scout becomes an increasingly capable "accessibility scout", tailoring accessibility scans to an individual's mobility level, preferences, and specific environmental interests through collaborative Human-AI assessments. We present findings from three studies: a formative study with six participants to inform the design of Accessibility Scout, a technical evaluation of 500 images of built environments, and a user study with 10 participants of varying mobility. Results from our technical evaluation and user study show that Accessibility Scout can generate personalized accessibility scans that extend beyond traditional ADA considerations. Finally, we conclude with a discussion on the implications of our work and future steps for building more scalable and personalized accessibility assessments of the physical world.

This work is licensed under a Creative Commons Attribution 4.0 International License.
UIST '25, Busan, Republic of Korea
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747624

CCS Concepts

• **Human-centered computing** → **Accessibility systems and tools**; **Interactive systems and tools**.

Keywords

Accessibility; Large Language Model; Accessibility Assessment; Personalization; Computer Vision

ACM Reference Format:
William Huang, Xia Su, Jon E. Froehlich, and Yang Zhang. 2025. Accessibility Scout: Personalized Accessibility Scans of Built Environments. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3746059.3747624

1 Introduction

Safe and accessible spaces are crucial for human well-being [19, 29, 39, 47] and quality of life [62]. However, these spaces are not guaranteed, especially for people with limited mobility who often cannot explore or use certain environments without proper accommodations. According to the CDC, 12.2% American adults have a mobility disability [14], with a majority of people expected to face mobility challenges as they age. In response, significant efforts have been made to identify accessibility concerns to renovate or build more accessible physical environments and inform people with limited mobility about potential challenges in unknown places [22, 25, 27, 63, 75, 82].

To scalably identify accessibility concerns, accessibility practitioners have codified common accessibility problems into standardized accessibility checklists like the Home Safety Self-Assessment Tool [33] and ADA building codes [3]. These checklists allow non-experts to evaluate and enforce the accessibility of environments more easily. Researchers have expanded upon this approach through automated accessibility auditing tools like RASSAR [74], which uses mobile AR and computer vision to identify pre-defined accessibility features and crowdsourcing platforms like Project Sidewalk [66], which defines a set of key sidewalk accessibility features for crowdworkers to annotate. While these systems are a cost effective and easily scalable way to collect large amounts of data, their checklist-based approach to identifying accesibility accessibility concerns fail to consider an individual's unique abilities, needs, preferences and how they change over time. Thus, the accessibility information generated from these approaches fail to capture how people personally experience accessibility in their physical environment [23, 40, 41, 45, 52, 57, 62, 76]. This mismatch can create a misleading perception of accessibility, potentially leading people with disabilities into frustrating and dangerous situations or imposing needless restrictions that further limit spatial opportunities.

In response, we propose a novel accessibility auditing approach leveraging recent advancements in large language models (LLMs), enabling personalized accessibility scans of built environments at scale. We present Accessibility Scout, an LLM-based personalized accessibility assessment system to semi-automatically identify accessibility concerns from images. Accessibility Scout accepts images of built environments to identify and visualize personalized accessibility concerns, enabling users to analyze thousands of environments at scale using data readily available from sites like Yelp or Google Maps. Through collaborative annotations, our system allows people with disabilities that affect their mobility to validate outputs from LLMs while incrementally learning their motor capabilities, preferences, and environmental interests to continuously updating its user model and improve its assessments over time.

We first ground our work by conducting a formative study to identify the current difficulties, needs, and process of finding accessible spaces. These insights informed the design of Accessibility Scout, aligning its prediction pipeline with how users naturally evaluate accessibility concerns and enabling users to guide the system through collaborative human-AI annotations to build a dynamic model of the user's needs and preferences. We then recruited 10 participants with varying levels of self-described mobility to build personalized user models through hour-long interactions with Accessibility Scout and conducted a technical evaluation using generated personalized accessibility annotations across 500 images of built environments. Finally, we conducted a user study where participants evaluated the usefulness of personalized versus non-personalized system outputs and shared their thoughts through interviews. These evaluations yielded both quantitative and qualitative insights about our system. Our findings indicate that Accessibility Scout effectively generates useful data and engages users, while raising important considerations for future AI systems for accessibility scans.

Our contributions are threefold: First, we introduce the first LLM-based approach to accessibility auditing, enabling semi-automatic, personalized assessments. We show that scalable personalization is possible using low-cost inference methods and readily available online data. Third, we provide novel insights from our user-centered design and evaluation, highlighting effective LLM user modeling strategies, new accessibility scanning methods, and new interaction techniques for human-AI collaborative annotations in accessibility. We believe Accessibility Scout paves a new path toward personalized accessibility assessment, with the potential to transform current accessibility practices. This potential is amplified by the vast number of online images (*e.g.*, from Airbnb to Yelp), which Accessibility Scout could use to support people with limited mobility in making travel decisions, room reservations, and event planning.

2 Related Work

We situate our work in digital accessibility assessment, visual affordance predictions, LLM-based personalization, and mobility modeling for accessibility.

2.1 Digital Accessibility Assessment

Recent innovations in sensing and computing have advanced environmental accessibility assessments. For example, Bring Environments to People [18] lets people with limited mobility remotely assess spaces through browser-based virtual tours. Other research [32, 38, 58, 60, 61] builds upon this idea by utilizing embodiment techniques in virtual reality (VR), allowing users to explore and assess digital twins of physical environments. Crowdsourced approaches like Project Sidewalk [66, 70, 77] use online crowdworkers to remotely label accessibility issues. Researchers have also attempted to automate accessibility auditing using computer vision to capture precise measurements and detect key issues in built environments using smartphones [74], robotic mechanisms [72, 73], existing imagery from Google Earth and Google Street View [36, 68], 3D scenes [28], scene graphs [24, 28], and point clouds [6, 8, 69]. However, current digital accessibility assessment solutions generally lack sufficient customizability and personalization features to support the diverse needs of the disability community. Accessibility

Scout addresses this issue by using LLMs and human-AI collaborations to continuously model a user's needs and generate more personalized accessibility scans.

2.2 Visual Affordance Prediction

Evaluating accessibility through images can be viewed as an application of visual affordance prediction, the process of using visual cues to identify how an object should be used. Recent research have shown how we can infer affordance from images through the innate properties of objects [15, 81], physical and social boundaries in a scene, [20], and specific interactions like grabbing [71]. Affordance prediction has become especially important with the rise of semi-autonomous robots, where robots must first identify whether a certain action is feasible before attempting it [11, 17, 56]. This process is similar to accessibility evaluations, where users must identify the feasibility of completing specific actions before traveling to the location. Closer to our work integrating recent developments in LLMs, recent developments in computer vision demonstrate how LLMs can be grounded to produce better affordance predictions [16, 64] and how LLMs can leverage visual affordance to improve outputs [17, 48]. We view Accessibility Scout as a human-centered approach to LLM-based visual affordance approaches where users leverage chain-of-thought prompting techniques to guide LLMs to consider specific tasks and that task's affordance in relation to the user's capabilities and environmental features.

2.3 LLM-based Personalization

Interest in LLM-based personalization is growing across various domains [51, 78, 85, 86]. Recent research has used LLMs to simulate human test subject responses in Turing tests [4] and replicate individual attitudes and behavior in interview responses [31, 59]. LLM agents have also been used to simulate user feedback on user interface usability [26, 79]. Harrak *et al.* [10] used LLM-based personalization to synthesize on-demand feedback from a target audience with LLM-based personas.

Closer to our work using LLMs to generate accessibility insights that influence how users choose environments are LLM-based personalized recommendation systems. Researchers have used LLMs to capture preferences of blind and low-vision individuals for navigational aid [5]. Joko *et al.* [37] demonstrated the use of LLMs to guide the creation of more aligned user preferences that better match users' actual preferences. Other research explored how LLMs can democratize personal health insights [21, 53] and create personal medical assistants [83]. LLMs have also been used to enhance traditional recommendation algorithms. Zhang *et al.* demonstrates the addition of smartphone sensory data to improve the emotional response of recommendations through LLMs. Other works demonstrate that the integration of LLMs into existing systems can directly improve recommendation quality [9, 49, 50].

2.4 Mobility Modeling for Accessibility

HCI reseachers have modeled motor capabilities to evaluate ergonomics [43, 44, 67] and assess the usability of different technologies and systems for a variety of audiences. While these user modeling frameworks were designed for general populations, other works focus primarily on people with limited mobility. Huang *et al.* [35] improves pose estimation for wheelchair users using synthetic data from photorealistic avatars driven by user-centered motion generation techniques. More closely related is literature on accessibility simulations using virtual human agents. Kaklanis *et al.* [38] models older adults and people with disabilities by breaking down tasks into hierachical motions and interactions for ergonomics testing of product prototypes in different scenarios and tasks. *Embodied Exploration* [60] models wheelchair users with avatars in VR using three key dimensional parameters: wheelchair maximum width, wheelchair armrest height, and seated eye height. These parameters are used to enhance the embodied experience, enabling more accurate accessibility assessments in VR environments. More automated utilization of user models include recent works in transportation network accessibility evaluations, using GIS-based networks [54, 84] to model how user preferences and capabilities might affect route finding.

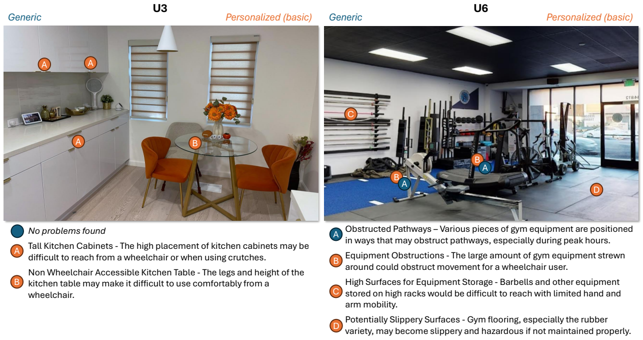

U3
Generic
Personalized (basic)
A
A
A
B
● No problems found
A Tall Kitchen Cabinets - The high placement of kitchen cabinets may be difficult to reach from a wheelchair or when using crutches.
B Non Wheelchair Accessible Kitchen Table - The legs and height of the kitchen table may make it difficult to use comfortably from a wheelchair.
U6
Generic
Personalized (basic)
C
B
A
B
A
D
A Obstructed Pathways – Various pieces of gym equipment are positioned in ways that may obstruct pathways, especially during peak hours.
B Equipment Obstructions - The large amount of gym equipment strewn around could obstruct movement for a wheelchair user.
C High Surfaces for Equipment Storage - Barbells and other equipment stored on high racks would be difficult to reach with limited hand and arm mobility.
D Potentially Slippery Surfaces - Gym flooring, especially the rubber variety, may become slippery and hazardous if not maintained properly.

Figure 2: Two examples that show the comparisons between generic LLM-generated annotations and those with basic personalization for two participants in our formative study.

Accessibility Scout introduces a dynamic new approach to digital accessibility assessment by harnessing recent advances in LLM-based collaborative annotation, visual affordance prediction, and personalization. Unlike traditional methods, Accessibility Scout empowers users to construct individualized user models that adapt to their specific needs and preferences, enabling image-based accessibility evaluations that more accurately reflect how they would experience and navigate a given environment before travel.

3 Formative Study

To inform the design of an LLM-based accessibility assessment system, we performed an initial formative study with the following goals: (1) to advance understanding of the current practices and challenges of accessibility assessment for people with limited mobility; (2) to investigate the feasibility of using LLMs for accessibility assessments; (3) to elicit user feedback on the idea of personalized accessibility scans. We recruited six participants (U1-U6), all self-identified as daily wheelchair users and were compensated with $20 for their time. Refer to Appendix Table A.1 for demographics.

3.1 Procedure and Analysis

Before the study, participants were sent an initial survey about demographics and self-described motor capabilities. Two researchers then conducted a ~30-min virtual research study with each participant individually through online meetings. Our study was approved by our institution's IRB.

Each study began with a semi-structured interview where participants were asked about their current experience virtually analyzing environments to further understand the needs and challenges of existing solutions. Participants were then shown images of different places and asked to think-aloud on how they would assess the accessibility of the environment. Researchers then generated two sets of accessibility concerns using OpenAI GPT-4o-2024-08-06 [1] by prompting *with* and *without* the user's self-described motor capabilities listed in Table A.1. Researchers then visually annotated each concern on the image and asked participants which set of annotations they preferred.

We collected audio recordings and notes, which were analyzed via thematic analysis [12] by two researchers to identify key themes and concerns. The first researcher, who also conducted the interviews, reviewed all transcripts and notes to develop an initial codebook. The two researchers then discussed the codebook, iteratively resolved disagreements, and developed a final version of the codebook. The first researcher then converted this codebook into identified themes. Participant quotes have been lightly edited for concision, grammar, and anonymity.

3.2 Findings

We highlight four key findings below in regards to existing accessibility assessment practices, personalization in accessibility, and reactions to LLMs for accessibility assessment:

Existing accessibility assessments deter exploring new environments. All participants agreed that current methods for environmental evaluation were difficult or insufficient. Many participants cited this as a reason why they do not explore new environments with comments like *"even though I've done my due diligence to ensure accessibility to my needs, there's a reasonable chance that I'll get there, and it's still going to **** up."* (U6). The tedious process of finding data can be a deterrent in and of itself, with U5 stating that *"I know we might want to go try a new place and with having to Google [accessibility]. I might just say, no. Let's just go somewhere where we know it's going to be accessible. So that is definitely a deterrent"* (U5).

Merits of LLMs for automatic accessibility assessments. All participants stated that the LLM annotations with and without their self-provided information were useful, especially when compared to existing accessibility data available online. Researchers also note that all participants found that the LLM's generated accessibility scans were equal or improved when prompted with the participant's self-described mobility. Example feedback regarding the usage of the LLM included *"It's amazing! It's not only wonderful information, it makes you feel more included too"* (U3); *"This is, gonna be so useful. It's going to be helpful even."* (U5); *"Because [the LLM] will identify restaurants or hotels that will address my needs"* (U4).

Need for personalized accessibility assessments. Throughout the study, participants appreciated the personalized accessibility assessments as seen in Figure 2. Researchers also noted that all participants evaluated environments differently during the think-aloud exercise, focusing on different aspects depending on their unique needs. During the annotations evaluations, U3 appreciated the personalized annotations mentioning their specific needs, stating *"It mentions that the navigating with wheelchair or crutches. That's beautiful. You don't see that a lot. Usually it just concentrates just on wheelchairs. So that is so awesome that that's mentioned."* (U3). U6 pointed out that his assessment of accessibility depends on the task at hand where *"The intended function. Is it going to be my permanent home? Is it going to be a temporary residence. Yeah, it absolutely changes my requirements of the space"* (U6).

Supportive features to improve usability. Researchers found that participants varied in the amount of detail on accessibility concerns they preferred. U1 points out that when they are conducting research on environments, *"I'm looking for a particular thing. If I wanted more maybe I could click to get more, but the main points are just simple and quick."* (U1). U5 requested more detail stating, *"There's different levels of wheelchair accessibility. So that extra detail is super helpful."* (U5). Participants also valued the use of visual markers to indicate accessibility concerns with comments like *"I could see how the [visual annotations] would be very useful, because [other people] are not in a wheelchair. They don't see those things that I do."* (U5). U6 also suggested that the system could be extended further to automatically handle environmental inquiries, *"I think you could do the calling in, if not with a you know actual voice, certainly by automated emails."* (U6).

4 System Design and Implementation

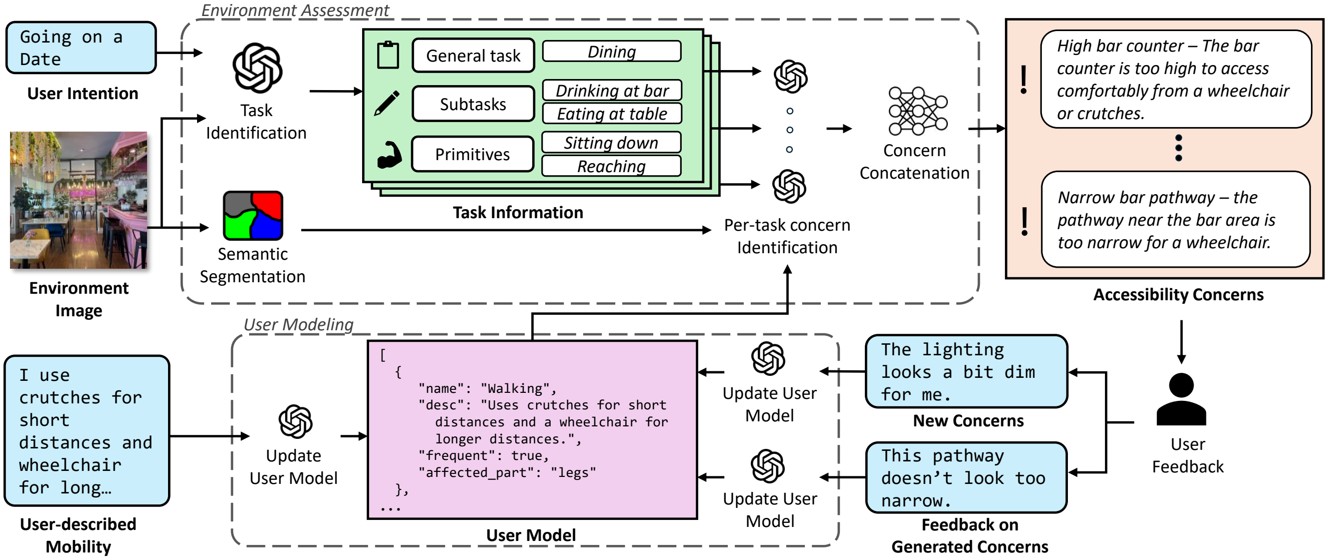
Following findings from our formative study showing how simple LLM prompts can produce useful accessibility scans, we developed Accessibility Scout, which combines LLMs and computer vision to model a user's accessibility preferences and enable semi-automatic, personalized, accessibility scans. All system components were developed using OpenAI ChatGPT-4o-2024-08-06 [1]. See Figure 3 for a system diagram.

4.1 Design Considerations

We developed Accessibility Scout with the primary objective of enabling users to run accessibility scans personalized to their individual needs at scale. We report on the following design considerations derived from our formative study:

D1: Support context-aware and adaptive personalization. Participants in our formative study often qualified their assessment of accessibility by the specific times and tasks they engage in. This goal aligns with findings from Lättman *et al.*'s [45], which detailed how perceived accessibility is highly context-dependent. Participants also emphasized that their accessibility needs change over time, whether due to physical decline or new techniques to navigate previous inaccessibilities. Therefore, a robust accessibility scanning system should support adaptive personalization by allowing users to provide feedback to address evolving needs and situational contexts.

D2: Enable human oversight over AI. While initial findings from our formative study indicate LLMs are capable of generating useful accessibility scans, we must acknowledge their susceptibility to inaccuracies and hallucinations. Participants echoed these sentiments, emphasizing the need for a tool that works *with* them and not *for* them. To support this, users should be able to review and



Going on a Date

High bar counter – The bar counter is too high to access comfortably from a wheelchair or crutches.

Narrow bar pathway – the pathway near the bar area is too narrow for a wheelchair.

I use crutches for short distances and wheelchair for long…

```
[
  {
    "name": "Walking",
    "desc": "Uses crutches for short
      distances and a wheelchair for
      longer distances.",
    "frequent": true,
    "affected_part": "legs"
  },
...
```

The lighting looks a bit dim for me.

This pathway doesn't look too narrow.

Figure 3: Accessibility Scout system structure. Users first submit an image of an environment and their intent. Accessibility Scout identifies potential tasks and their actions to predict environment concerns. The user can then provide feedback on these concerns to update their user model, which leads to improved accessibility assessment in future scans.

revise LLM-generated annotations, facilitating both personalization and error correction. Our system should also be easily interpretable, helping users to easily understand and guide the scanning process.

D3: Generate detailed accessibility concern information. Participants preferred varying levels of detail in accessibility information depending on how important it was to access a given space. They also highlighted the value of being able to explore additional details about a concern when needed. Accessibility Scout should support these needs by providing a variety of accessibility information, allowing users to quickly glance at key insights or conduct a more in-depth analysis of the built environment as needed.

D4: Enable the system to run at scale. Participants in our formative study were especially excited about using a potential LLM-based system to more efficiently discover accessible environments. For example, participants envisioned using such a system to evaluate multiple Airbnbs more efficiently than would be possible through manual inspection, simplifying the search for accessible vacation options. To support such use cases, Accessibility Scout must be able to scale to handle large volumes of built environments.

4.2 User Modeling

Accessibility Scout's user modeling component handles the maintenance of a structured user model which can be iteratively updated through user feedback and is easily interpreted by LLMs.

User model structure. To facilitate both human interpretability and AI performance, we represent the user model in JavaScript Object Notation (JSON) format. Each user model consists of a set of attributes which describes a specific movement (*e.g.*, reaching above with my right arm), how the movement might be affected (*e.g.*, I can not reach above shoulder level), whether the movement is frequently performed, and the affected body part or preference (arms, legs, feet back, chest, hands, eyes, ears, brain, user preference). An example of a user model attribute is shown in Figure 3. This user model structure can capture physical attributes of a user (*e.g.*, footprint of a wheelchair), sensory and cognitive attributes (*e.g.*, sensitivity to sound), and the user's value system (*e.g.*, prefer quieter places) which has been shown to better represent a user's decision making [62], and provides additional context to the importance of specific subtasks through the frequency boolean. Furthermore, JSON attributes allow various details, including context and specific scenarios, which can grow and shrink over time.

Elicitation methods. To generate the user model, we enable three different elicitation methods: (1) *Self-Description.* Users are able to enter an unstructured textual description of their capabilities and preferences which can include recounts of prior experiences. An LLM then decomposes this self description into a series of affected motions which can be input as the user model. (2) *Environmental Annotations.* Users can also choose to annotate concerns in images of different environments. The concerns, their reasoning, and the image are entered into an LLM to generate a user model. (3) *Feedback on Environmental Annotations.* Users can also update their user model by providing feedback to AI-generated environmental annotations. User feedback, the original annotation, and the image are then inputted into an LLM and used to update the user model. All elicitation methods are implemented through textual input. User interfaces to utilize elicitation methods are detailed in Section 4.3.

4.3 Accessibility Assessment

Accessibility Scout uses the user model to generate accessibility concerns on images of different environments (Figure 3). In order to build a more human-interpretable and controllable system, we designed the assessment process to mimic how participants from our formative study evaluated the accessibility of places by the tasks they may prohibit. This serves three primary purposes: (1) To

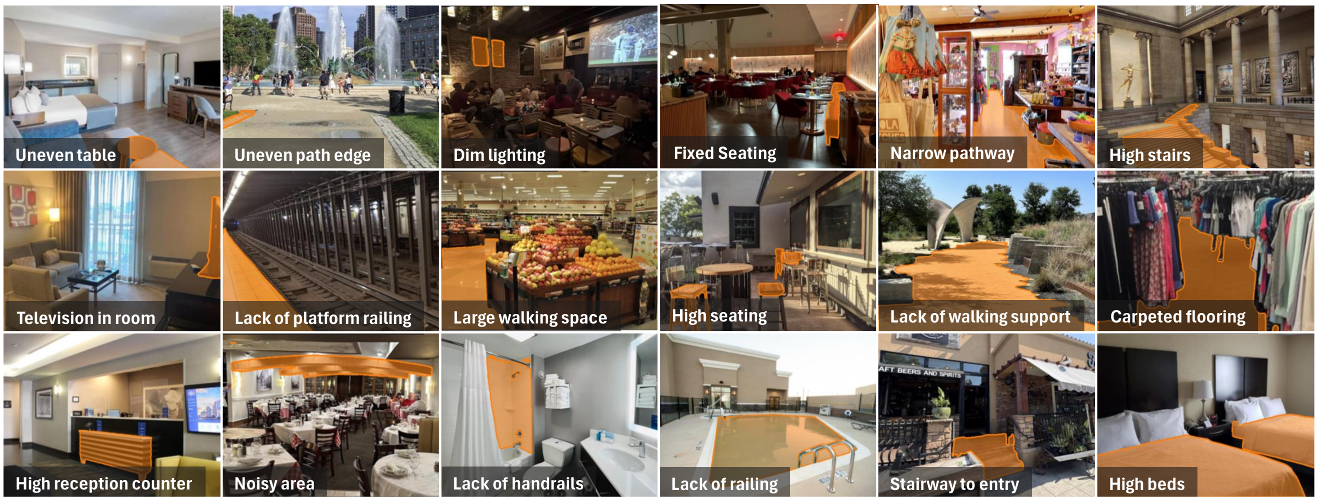


Figure 4: Example accessibility concerns identified by Accessibility Scout. These concerns range from narrow floorplans and furniture height to general facets of environment like the presence of specific objects or noise. Note: texts shown are summarizations from detailed concerns. Full unabbreviated examples are in Figure 1.

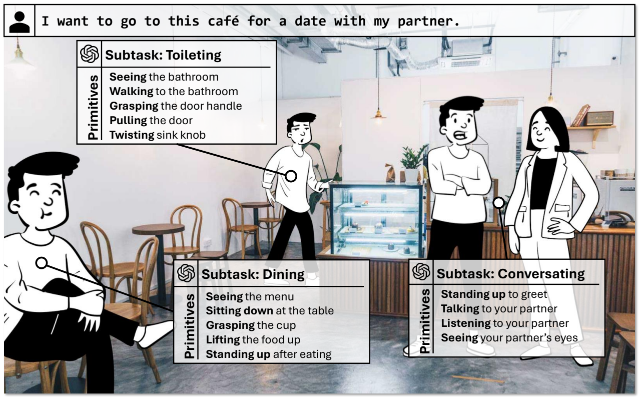
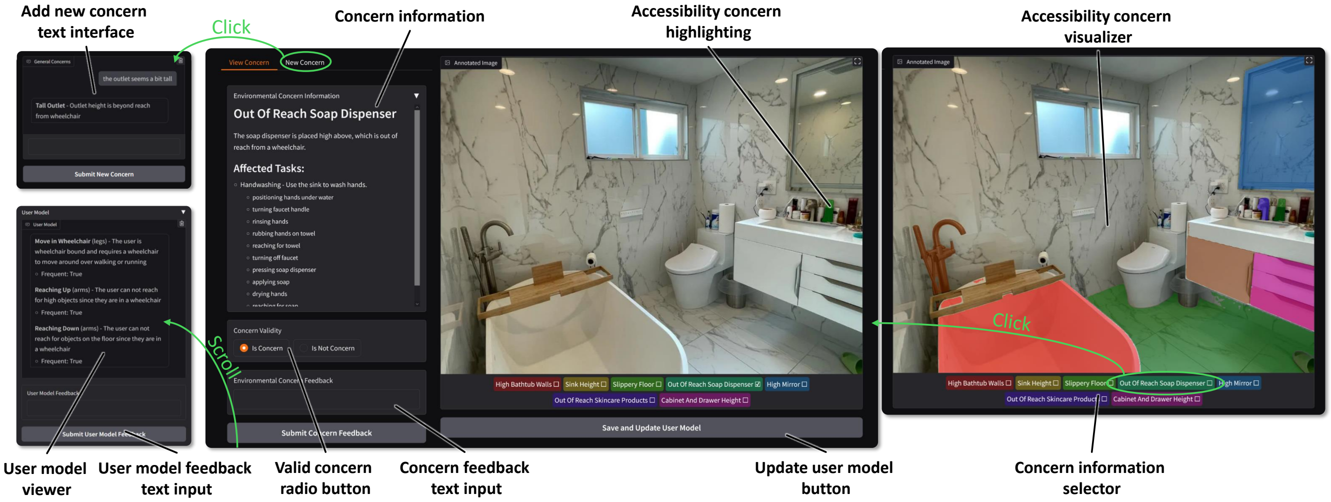


Figure 5: Visualization of how tasks are broken down by Accessibility Scout. Given a desired user intent like "Going on a date", the system first breaks down the task into subtasks the user might perform during the task like "toileting", "conversating", or "dining" and potential locations for these subtasks. Upon doing so, the system then further breaks down these subtasks into primitive motions.

address the need for task-specific accessibility assessments identified through our formative study. (2) To generate more relevant and interesting concerns. (3) To enable parallelized LLM requests for scalability and reliability.

Task identification. To generate comprehensive and detailed accessibility scans, we first identify the spatial tasks a user could engage in based on their intended use of the environment. Users first input an image of an environment and a short description of the environment and their intended usage into an LLM which is prompted to predict a set of common tasks that a user might perform in the environment (*e.g.*, study at a cafe). Within the same context window, the LLM is then prompted to decompose each one of these tasks into subtasks (*e.g.* reading your textbook is a subtask of studying at a cafe). Each subtask consists of a short description, potential locations these subtasks might be performed in an environment, and primitive motions necessary, a concept derived from Kaklanis et al.[38] which models any task as a series of fundamental movements like grabbing, reaching, and pulling. Figure 5 illustrates an example of the identified tasks and fundamental motions required when a user specifies their intended use of the environment for a date. By first breaking down the environment into potential tasks and subtasks, we treat accessibility assessments as a task affordance problem which is both more reflective to what a user might actually need and enables more focused LLM contexts to generate better predictions later on. Prompts for are shown in Appendix Figure B.1 and Figure B.2.

Concern identification. To identify key environmental concerns, we use Set-of-Mark Prompting [80]. Images are overlaid with semantic segmentation masks generated from Semantic-SAM [46], providing textual labels for different semantic segmentation masks of the image to enhance the LLM's spatial understanding. Masks are later used for visualization purposes in the user interface. For each task generated from the task identification process, its task description, list of subtasks, the user model, and Set-of-Mark prompts are fed into an LLM prompted to identify key parts of the environment that would prohibit the task. For each task, the LLM outputs a set of environmental concerns. Each concern consists of a short name, reason for why the concern was identified for the user, and the location by the label generated from Semantic-SAM segmentation. We note these concerns are qualitative heuristics like *"low"* or *"soft"* and not precise measurements. Each task is processed in a parallel LLM request for speed, smaller context and more focused context windows, and partial outputs in cases where the API endpoint is unstable. Prompts for this process are shown in Appendix Figure B.3.


Add new concern text interface
Click
Concern information
Accessibility concern highlighting
Accessibility concern visualizer
General Concerns
Submit New Concern
View Concern
New Concern
Annotated Image
User Model
Scroll
Concern Validity
● Is Concern
○ Is Not Concern
Environmental Concern Feedback
Submit Concern Feedback
User Model Feedback
Submit User Model Feedback
High Bathtub Walls □
Sink Height □
Slippery Floor □
Out Of Reach Soap Dispenser ☑
High Mirror □
Out Of Reach Skincare Products □
Cabinet And Drawer Height □
Save and Update User Model
Click
Out Of Reach Soap Dispenser □
User model viewer
User model feedback text input
Valid concern radio button
Concern feedback text input
Update user model button
Concern information selector

the outlet seems a bit tall

Tall Outlet - Outlet height is beyond reach from wheelchair

Environmental Concern Information

Out Of Reach Soap Dispenser

The soap dispenser is placed high above, which is out of reach from a wheelchair.

Affected Tasks:

- Handwashing - Use the sink to wash hands.
  - positioning hands under water
  - turning faucet handle
  - rinsing hands
  - rubbing hands on towel
  - reaching for towel
  - turning off faucet
  - pressing soap dispenser
  - applying soap
  - drying hands
  - reaching for soap

Move in Wheelchair (legs) - The user is wheelchair bound and requires a wheelchair to move around over walking or running
- Frequent: True

Reaching Up (arms) - The user can not reach for high objects since they are in a wheelchair
- Frequent: True

Reaching Down (arms) - The user can not reach for objects on the floor since they are in a wheelchair
- Frequent: True

Figure 6: The Accessibility Scout UI. Right: The user can view all detected concerns in the visualizer. Middle: Upon selecting a specific concern, information on the concern and highlighting are then shown. The user is then able to provide feedback on the concern to improve the user model. Top left: The text interface for users to add more concerns not identified already. Bottom left: The text interface for users to view their user model and provide unstructured text feedback.

Concern concatenation. Since concerns are generated in parallel, requests can generate redundant accessibility concerns. Thus, all concerns are grouped through semantic text similarity analysis using Sentence-BERT `all-MiniLM-L6-v2` [65] with the name and reason of each concern. Concerns with a cosine similarity over a threshold of 0.7, selected through experimental analysis, are combined by selecting the name and reasoning with the highest average cosine similarity in the group (*e.g.*, *High Bar Counter - The bar counter is too high for the user to access comfortably from a wheelchair* and *High Bar Counter - The height of the bar counter makes it difficult for the user to reach drinks or interact comfortably*). Examples of final accessibility concerns are illustrated in Figure 1, highlighting various issues identified for users with different mobility levels and preferences. Figure 4 presents a broader range of compatible photos of built environments.

4.4 User Interfaces

Accessibility Scout has a web user interface implemented in Gradio [2] that takes images of environments and generates indicators on potential accessibility concerns which users can provide feedback on (Figure 6). The following describes an example user scenario:

Alexandria is a wheelchair user using Accessibility Scout to evaluate the accessibility of an Airbnb. Upon starting Accessibility Scout with a picture of the Airbnb's bathroom and her user model, Alexandria is greeted with a visualization of all detected concerns (Figure 6 right). She hovers her cursor over the selector labeled *Out of Reach Soap Dispenser* which highlights the concern in the visualizer. She then clicks on the selector which changes the accessibility information textbox to show that the soap dispenser is inaccessible because it is too high on the counter (Figure 6 middle). Alexandria agrees with this generated concern, selecting the *Is Concern* radio button, and provides further feedback in the *Envrionmental Concern Feedback* textbox that *"The soap dispenser is just too far back on the counter to reach comfortably"*. Alexandria repeats this process for all other identified concerns before noticing that she might not be able to plug in a hairdryer since the outlet is high up. She clicks the *New Concern* button and types in the text interface: *"The outlet seems a bit tall"* (Figure 6 top left). She presses enter (same as clicking on the *Submit New Concern* button) and the text interface and visualizer show a new concern: *Tall Outlet*. She scrolls down to the user model viewer and sees a new attribute: *Outlet height is beyond reach from wheelchair* (Figure 6 bottom left). She finally finishes evaluating all generated concerns and clicks *Save and Update User Model* which prompts Accessibility Scout to update her user model with her feedback on this image.

5 Technical Validation

We investigate the feasibility of using LLMs to assess the personalized accessibility of environments and ground our findings from the subsequent user study through a set of technical evaluations of Accessibility Scout in 500 images of different environments using varying user models. Through our technical evaluations, we note that it is unfeasible to directly evaluate the quality of identified concern heuristics given their accuracy is subjective (*e.g.*, a table of 34 inches might be okay for someone in a manual wheelchair to sit at but too low for a power wheelchair). Instead, we defer evaluations on usefulness to our subsequent user study. In this section, we focus on evaluating the following properties of Accessibility Scout: (1) Accurate detection of environmental features as a measure of hallucinations. (2) Distribution of identified concerns as a measure of Accessibility Scout's ability to capture accessibility needs in different environments. (3) Differences in detected concerns between user models as a measure of degree of personalization. (4) Cost of evaluation as a measure of scalability.

5.1 Data Collection

Researchers compiled 500 images of different environments across 9 of the most populous cities in each region of the United States (Los Angeles, San Diego New York City, Philadelphia, Chicago, Columbus, Phoenix, San Antonio, Houston) sourced from searches through Google Maps and Yelp. Eight crowdworkers first used keywords of commonly accessed environments (community centers, grocery stores, lodging, restaurants, retail stores, transportation hubs, and public venues) across North America to assemble an initial dataset of images of environments. Researchers then manually evaluated and selected 500 images according to the following criteria: Images showed key parts of the environment including pathways, utilities, restrooms, and functional areas necessary to complete the purpose of the area. Images had a wide enough field of view to capture the entire environment (*e.g.* full view of the floor up). All images were then briefly labeled with a general description of what someone might be doing in that environment. Examples from the dataset are shown in Figure 4. Using Accessibility Scout, we run an accessibility scan of each image using each of the user models created by participants in the first stage of a later user study, which will be detailed in Section 5.5, as well as an empty *"generic"* user model for a total of 11 user models. Demographic information used to initialize Accessibility Scout are shown in Table 1. As a result, we generate 5,500 accessibility scans (39,394 concerns).

5.2 Detection Performance

Accurate and robust detection forms the foundation of accessibility assessment. We measure the number of misdectections using a fact-checking approach [34] which only measures object detection accuracy for detected concerns. Since accessibility is highly subjective, we evaluated our system's ability to find useful concerns entirely through our subsequent user study. Human evaluators manually reviewed all 500 accessibility scans generated by the generic user model, determining whether each identified concern was a hallucination based solely on the following criteria: (1) *Does the related concern exist in the image?* (2) *Does the concern correctly identify the object of concern?* We evaluate purely on this criteria and do not attempt to label object qualities like *"too high"* or *"too soft"* as true or false given users can perceive these qualities differently. Evaluators rated 3590 concerns and found 237 (6.63%) hallucinated concerns, which were removed from our later user study. Evaluators noted that hallucinated concerns often centered around specific environmental features not depicted in the image like *checkout counters* or *TV remotes*, potentially indicating that OpenAI-GPT4o has an existing bias towards specific environmental features.

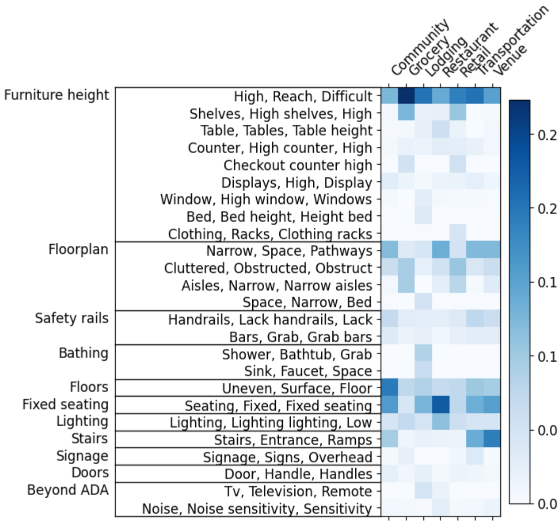


Figure 7: Distribution of twenty-five largest unsupervised clusters of concerns expressed as a percentage of total concerns in each environment type grouped by ADA categories. Labels are three representative keywords for a cluster generated by taking the three terms with the highest *Term Frequency-Inverse Document Frequency*, a measure of a term's importance within a group of text.

Table 1: Demographic and self-described mobility of 10 participants (P1-P10) in the user study.

ID	Age	Gender	Diagnosed disability	Self-described mobility and preferences
P1	58	F	Spinal cord injury, Fibromyalgia, Stenosis, Arthritis	Use a manual wheelchair with attached motor. Able to stand and pivot for transfers. Upper extremities are weak with limited strength and dexterity. Visual and hearing difficulties.
P2	54	M	Spinal cord injury (T11/T12, L4/L5)	Crutches for short distance and wheelchair for long distances. Not injured from waist up and have full function in chest, arms, hands, shoulders, and neck. Nerve damage in hips, hamstring, and no function from knee down.
P3	61	F	Polio, Post-polio syndrome	Ambulation causes physical and mental strain due to tripping risk. Stairs, inclines, uneven surfaces, and slick surfaces are difficult. Cannot stoop, rise off floor, or rise from low toilet without support. Stepping in/out of tubs is dangerous but possible with grab bars. High bar stools are difficult. Holding anything in one hand can be difficult as I walk with a forearm crutch. Long distances require a rollator. Muscle fatigue requires frequent breaks.
P4	72	F	No diagnosed disability	Cannot walk extended distances due to fatigue. Long distances, staircases are problematic. Handrails are important.
P5	86	F	Post-polio syndrome	Need to use a walker. Right leg is super weak, especially knee. Have a hyper extended knee and drop foot. Use a knee brace and ankle foot orthosis. Cannot walk long distances.
P6	35	M	Spinal cord injury (T10)	Use a manual wheelchair. Prefer rolling on hard floors. Carpet is difficult to push. Can move pretty comfortably and smoothly in wheelchair and can go anywhere except sandy places.
P7	58	M	Quadriplegic (C7)	Use a manual wheelchair and SmartDrive assistive device. Prefer hard floors, wide corridors, low/no thresholds in doorways, automated entry doors. Need restroom facilities that accommodate wheelchairs with designs that follow ADA. Appreciate buildings with elevators that I can operate over special lifts that require assistance. Appreciate using main entrances over special side entrances.
P8	38	M	Quadriplegic (C4)	Use an electric wheelchair. Wide hallways and doorways are a must. Prefer hard, smooth floors and no thick carpet. Need elevators with wide automatic doors and enough space inside. Ramps need a gentle slope. Need accessible restrooms with enough space to turn. I do not transfer so toileting and grab bars are not important. Prefer lever-style door handles or automatic door handles since twisting doorknobs is challenging. Paddle switches work best for lighting. Good lighting is a must to see everything. Need to watch for rough or uneven surfaces. Long distances are difficult due to battery constraints.
P9	66	M	Paraplegic (T4)	Use a manual wheelchair. Need doors wide enough for wheelchair. Prefer 1 level places and ramps/elevators instead of stairs. Prefer grab bars in rest rooms, hand drying equipment next to sink, hard floors, and doors that swing out.
P10	50	M	Quadriplegic	Use a power wheelchair and prefer environments with smooth flat surfaces and open. spaces.

5.3 System Cost

We also conducted a basic evaluation on the cost and scanning time of Accessibility Scout by measuring the average token usage to evaluate an environment. We compute an average token usage across 500 images of 8758 tokens/image (STD = 1112.175), 9 requests/image, and average delay of 10.737s (STD = 6.987). We therefore estimate the cost of using Accessibility Scout as $.021/image with the ability of running up to 3553 images/minute using ChatGPT-4o-2024-08-06 as of March 2025[1]. Researchers note that the time of day can greatly affect the scanning time and reliability as running Accessibility Scout during North American working hours would be drastically slower and lead to more dropped API requests. [1]

5.4 Distribution of Generated Concerns

To understand the distribution of these generated concerns, we conducted an unsupervised topic clustering of generated concerns using BERTopic [30]. The top 25 largest clusters were included for further categorization. First, researchers identified a set of key accessibility clusters related to ADA guidelines. Researchers manually assigned these clusters to ADA categories (*e.g.*, Furniture Height, Floorplan), and assigned the rest of the clusters to "Beyond ADA". This process resulted in 11 categories shown in Figure 7. We find that concerns tend to correlate with types of environments. For example, "Fixed seating" consists of a higher percentage of restaurant concerns (22.70%) and "High, Reach, Difficult" concerns are highly prevalent in grocery stores (27.31%) due to high shelves of items. We found two clusters that went beyond existing ADA categories: "tv, television, remote" and "noise, noise sensitivity, sensitivity". This result indicates the potential for Accessibility Scout to extend beyond ADA classifications of accessibility to individual needs of the user. Furthermore, the identification of a "noise" related cluster indicates that Accessibility Scout considers *people* surrounding a user with limited mobility as an important yet often overlooked facet, and how environments might change even if "noise" is not directly depicted. Our findings indicate that Accessibility Scout aligns with common-sense expectations of where accessibility concerns typically arise, demonstrating how our approach can generate high-quality accessibility data. Moreover, Accessibility Scout goes beyond detecting only visible environmental features, enabling dynamic and nuanced spatial inferences.

5.5 Accessibility Scan Personalization

To evaluate Accessibility Scout's personalization, we use the same clustering procedure as the previous section with a new analysis to compute the distribution of accessibility scan clusters across participants (Figure 8 left). Additionally, we measure the *Wasserstein Distance* between each participant's distribution scaled by the total number of concerns in each category which can be interpreted as the amount of work to transform one participant's concern distribution to another (Figure 8 right). We note that this metric does not reflect the quality of personalization, which is evaluated in the user study, but rather highlights the variation in accessibility predictions across participants. High Wasserstein distances align with differences in mobility across participants. Notably, P4 and P5 demonstrated the highest average Wasserstein distance (0.0708 and 0.0523) as the only participant who did not have a diagnosed disability and the only participant who walks with the assistance of a walker, respectively. Even among participants with low Wasserstein averages, we still note that there are noticeable differences in their distribution of concerns as shown in Figure 8 left. For instance, P6, P9, and P10 had a lower average Wasserstein distance (0.0262, 0.0286, and 0.0304 respectively) as P6 and P9 were both highly independent and believed that they could handle most challenges that came their way, often marking newly identified concerns as irrelevant during training, and P10 who believed that most concerns were not relevant as they were not able to access that environment feature in the first place given their highly limited mobility (*e.g.*, fixed seating was not relevant since they could not transfer at all). These results indicate that Accessibility Scout can differentiate between the unique needs of individual users to generate meaningfully different accessibility scans.

6 User Study

We conducted a final user study to better understand the capabilities of Accessibility Scout for personalized accessibility scans and the usability of the system by comparing the usefulness between concerns generated from a generic and personalized model. We believe a generic user model is analogous to static one-size-fits-all approaches to accessibility assessments like checklists while equally susceptible to hallucinations as the personalized model which allows us to evaluate the impacts of personalization on usefulness. We recruited 10 participants (P1-P10) with varying levels of self-described mobility. Participant demographic information is listed in Table 1. Participants received $100 compensation for completing the full study, which comprised two stages of approximately one hour each. Our study was approved by our institution's IRB.

6.1 Procedure

All participants were sent an initial survey requesting basic demographic information and a self-description of their physical and mental capabilities. Their self-description was then used to generate an initial user model. A dataset of images of different environments was then compiled through the following procedure. Participants were first asked to provide 15 different locations they have physically explored to help participants draw from previous experiences and better evaluate the performance of Accessibility Scout. For all evaluations, users were asked to take into account any prior knowledge they had visiting the depicted location, implicitly evaluating accuracy as well. Researchers then randomly selected 15 unfamiliar locations. For each location, researchers sourced one image following the same criteria as the technical evaluation data in Section 5.1. Two researchers then conducted the two-stage virtual user study with each participant through online meetings.

Each study began with an initial one hour stage where participants were asked to train Accessibility Scout through a user-guided accessibility scan process. Participants were first given an introduction and briefing on how to use Accessibility Scout to evaluate generated concerns and provide their feedback. Participants were given the option of manually controlling the system through remote desktop control or dictating actions to the researchers to use

[1] Pricing and timing estimates can vary greatly depending on selected LLM. Other LLMs can be used to reduce cost.

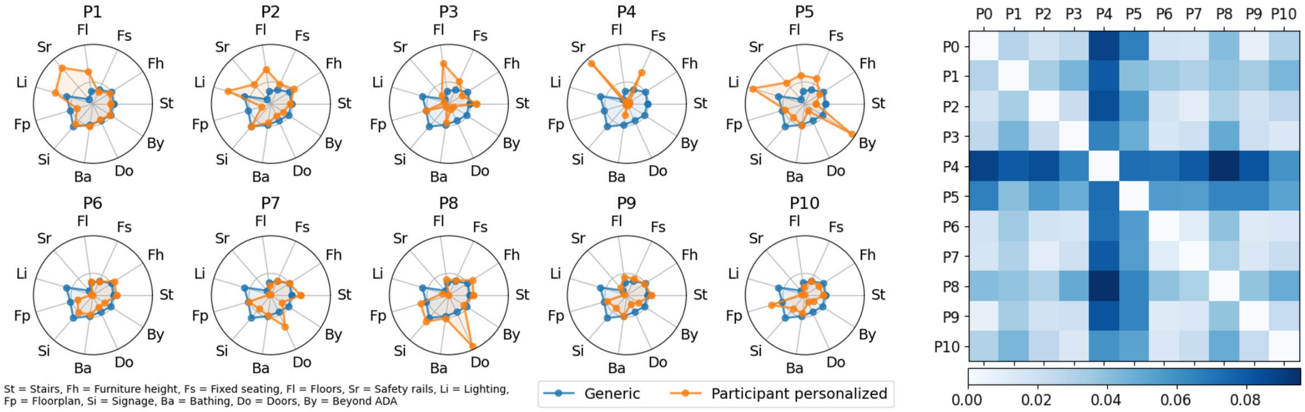
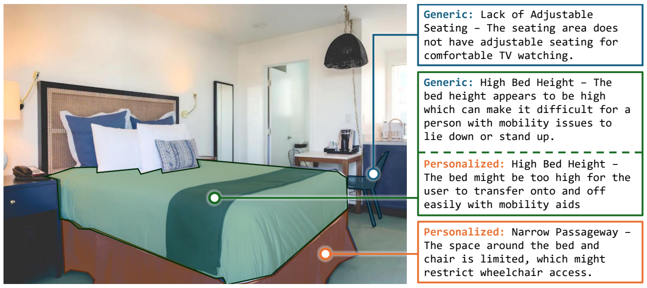


Figure 8: Left: Distribution of participant's generated concerns grouped by ADA categories. Differences in distribution shape indicate differences in Accessibility Scout's generation given a specific participant's user model. All data is displayed on a symlog scale. Right: Wasserstein distance between participant's generated accessibility scans. Higher value indicates greater difference between concerns. P0 indicates a "generic" user model or empty JSON, which was not trained by a real user in our study.


Generic: Lack of Adjustable Seating - The seating area does not have adjustable seating for comfortable TV watching.

Generic: High Bed Height - The bed height appears to be high which can make it difficult for a person with mobility issues to lie down or stand up.

Personalized: High Bed Height - The bed might be too high for the user to transfer onto and off easily with mobility aids

Personalized: Narrow Passageway - The space around the bed and chair is limited, which might restrict wheelchair access.

Figure 9: An example of the user study setup for P2. Users are asked to rate the usefulness of *Unique Concerns* (blue and orange) and *Similar Concerns* (green) against each other in a blind test.

the system. A random selection of 15 images from the compiled dataset was used in this stage.

Following the first stage of the user study where user's created their own user model, researchers conducted a blind test to analyze how well Accessibility Scout personalized to the user. Researchers scanned the remaining 15 images twice using Accessibility Scout, once with the trained user model and once with an empty JSON (generic) user model only representing the LLM's innate knowledge of accessibility. Participants were not informed of how the concern was generated (*i.e.*, from the personalized or generic model) until after the study concluded, ensuring a blind test process to eliminate potential bias.

First, 15 *Unique Concerns* (concerns present from one model's prediction but not the other) from each model were randomly selected and shuffled to analyze how well Accessibility Scout could identify personalized accessibility needs. An example of different unique concerns is shown in orange and blue in Figure 9. Participants were then asked to rate the usefulness of knowing each concern before visiting the environment on a 7-point Likert scale.

In addition, 15 *Similar Concerns* (concerns present in accessibility scans from both user models differing only in wording) were randomly ordered and visualized side by side to evaluate Accessibility Scout's personalization of concern descriptions. An example of similar concerns is depicted in green in Figure 9. Participants were then asked to evaluate the usefulness of knowing the concern before visiting the environment for both shown concerns on a 7-point Likert scale and which concern description they preferred.

After completing all evaluations, participants were informed that concerns were generated from Accessibility Scout and which concerns came from the generic vs. personalized user models. They were then interviewed to better understand any Likert scale ratings and discrepancies between the two user models. Participants were also asked to reflect on their overall experience, concerns about AI in accessibility, and the importance of personalization. These interviews were conducted in a semi-structured format.

Three researchers then conducted a codebook thematic analysis [13], a middle ground approach between structured and reflexive methods, using study recordings, transcriptions, and researcher notes. Researchers first converged on a set of a priori themes from prior research and the needfinding study. The first researcher, who also served as the interviewer, then inductively developed the codebook by reviewing all transcripts and notes in relation to the original themes. The three researchers then collaboratively refined the codebook, resolving disagreements and grouping codes into themes. Our thematic codebook is shown in Appendix Table C.2.

6.2 Findings

All ten participants were able to train and evaluate Accessibility Scout. A Wilcoxon Signed Rank Test, a rank-based nonparametric test, was then conducted to assess significant differences between the generic and personalized ratings within each user and across all users. We note that evaluating statistical differences within users can capture more insight into individual user variances at the risk of inflating Type I statistical errors. Figure 10 shows participants'

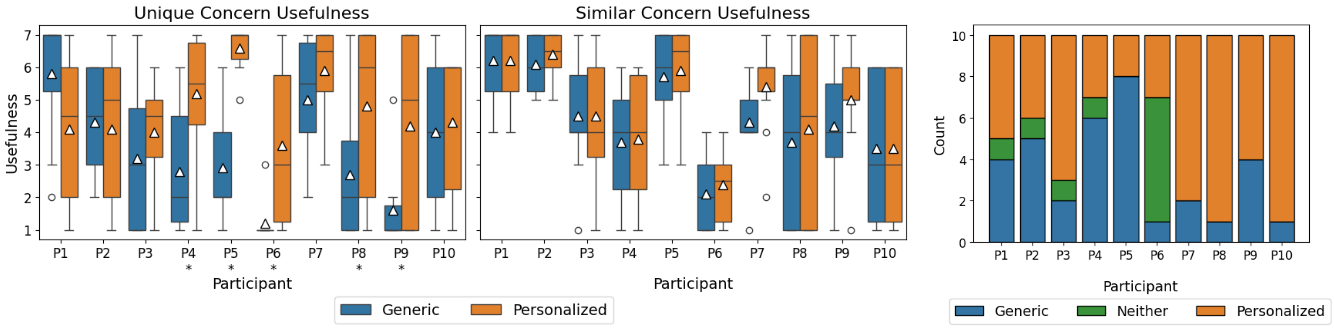


Figure 10: Left: 7-Point Likert Scale scores across the concerns unique to the generic or personalized user model and across the concerns that are present in both the generic and personalized user model. Large white triangles in the box denote means. "1" indicates that the concern was not useful at all and "7" indicates that the concern was very useful. "*" denotes $P < 0.05$.. Right: Preferences between generic and personalized models in similar concerns.

ratings and preferences. Across both studies, participants, rating the personalized accessibility scans as relatively useful (mean = 4.7/7, STD = 2.17).

Through our thematic analysis, we identified 51 codes (Table C.2). Some of the largest codes included "Accuracy and Detail of Concern" which highlighted improvements to the accuracy of our system, "Context Influences Accessibility Concerns" which captures how users also consider accessibility in regards to when and where something is (*e.g.*, winter vs. summer), and "AI Usage is Appropriate for Accessibility Scans" which discusses how users feel comfortable with AI for accessibility assessments. Below, we present key findings from our thematic analysis. Participant quotes have been lightly edited for concision, grammar, and anonymity.

Collaborative AI training is effective. As was demonstrated in our technical evaluations (Section 5.4), we found that generated user models from only one hour Accessibility Scout's collaborative training was effective in differentiating generated concerns. P2 believed that the current amount of personalization was perfect: *"even with the personalization we just put in, I think is great. I don't think it has to go further than that"* (P2). Furthermore, participants also enjoyed the process of training: *"Oh this is so much fun, I'm loving this"* (P2), *"This is fun"* (P5). At the same time, participants also noted that the collaborative design of Accessibility Scout helped them trust the system more, alleviating concerns on AI accuracy. P8 stated that being able to "double check" the AI was vital: *"when we went through the initial study and we were able to add my input into it, that was amazing like that was absolutely amazing. And you could see the changes from that...I think that's amazing. and I think that's very important"* (P8). P3 shared this sentiment, stating that control was vital for them to use an AI system: *"I'm in control. And in the end I can choose it or not, use it more, use it less, so in the end as long as I have control [over the AI]"* (P3).

Personalization generally makes accessibility scans more useful. Unique concerns generated from the personalized model were perceived as more useful than those from the generic (mean = 4.68, STD = 2.26, mean = 3.35, STD = 2.24, respectively ($p < .001$)), indicating that the addition of personalization generates more useful accessibility scans. P8 echoed these findings, stating that the accessibility scans from the generic user model were laughable: *"[the generic concerns], I would look at that and I would laugh and I would not look at it again you know. But the personalized information that was coming up like I found a lot of that information very useful"* (P8). P10 also found that the addition of personalization made the system more usable by filtering out unnecessary information: *"It would just make it easier if it was more personalized. It'd be less information to filter out...it would make it simpler and more efficient to have it personalized"* (P10). While personalization was generally perceived as useful, P4 notes that personalization can actually reduce the usability of Accessibility Scout when trying to plan for groups: *"[Personalization is] low importance...having more information also allows me to know that if somebody's coming to visit what I'm looking for...It lets me know for more than just myself"* (P4).

Mixed feedback on referencing user capabilities. Furthermore, while personalized accessibility scans were generally perceived as useful, the language used to convey this data had more mixed feedback. In comparing descriptions of the same concern, participants rated the personalized and generic model relatively similarly ($\text{mean} = 4.72/7, \text{STD} = 2.07$, $\text{mean} = 4.4, \text{STD} = 2.04$ respectively ($p = .187$)). We also find that participants only slightly preferred the wording of the personalized concern when compared side by side ($\text{mean} = 5.6/10, \text{STD} = 2.458$) (Figure 10 right). While viewing a concern which they regarded as not specific enough, P8 believed that generic descriptions are less useful: *"Makes [the concern] just a little bit less [useful] when it's so generic"* (P8). P5 noted that mentioning their capabilities helped draw their attention to key accessibility concerns: "To me [the generic and personalized descriptions] look very similar. But the [mentioning of a] walker is just a red light to me, or an alarm bell" (P5). When evaluating a concern about staircases causing fatigue, P4 notes that the personalization of Accessibility Scout was drawing conclusions for them: *"Having it say could cause fatigue just thinking out loud seems overly narrow and irrelevant...you don't draw the conclusion for me"* (P4). After reading an explanation for how stairs were not accessible since they used a walker, P3 believed that extra language personalizing the accessibility scans obfuscated the important data: *"I don't need to read whole paragraphs about things. I just need [to see it]* *and it kind of points [concerns] out"* (P3). These findings suggest that the representation of accessibility concerns is not as important as identifying them in the first place. This is further compounded as participants appreciated the ability to verify identify concerns, rendering the concern descriptions redundant in many cases.

Potential applications. During our user study, participants offered a variety of different use cases for Accessibility Scout. P7 viewed Accessibility Scout like a pre-game report to lower the uncertainty in new situations: *"The more I can get advanced scout reports [on accessibility information], the more I can avoid all the uncertainty and angst of the first couple of visits"* (P7). P9 shared similar sentiments, stating that Accessibility Scout could help them plan: *"If I know more about the location then I can bring things with me to help overcome the deficits that I do have"* (P9). While P6 believed the system was not as useful for scouting unavoidable environments, they found utility in using Accessibility Scout to find places to go to.

System improvements. During our user study, participants shared detailed suggestions for improving Accessibility Scout. All participants believed that the explanation for why a concern was found could be more accurate to their individual needs and the environment. P1, P3, and P7 suggested that the concern descriptions could be made more concise. P8, P9, and P10 noted that the visual highlights were sometimes inaccurate and distracted the user. P3 and P4 believed that the ability for users to see human feedback on the environment would help them trust the system more. P8 and P10 also noted that Accessibility Scout would sometimes show irrelevant concerns for tasks they could not complete at all (*e.g.* Accessibility Scout showed inaccessible seating when they are unable to transfer seating at all). Finally, researchers also noted that concerns would occasionally duplicate, signaling a need to improve concern concatentation in future versions of our system.

7 Discussion

We discuss key implications of our findings, limitations, and opportunities for future work.

Application scenarios. Accessibility Scout's LLM-based approach supports highly personalized and scalable accessibility auditing, appealing for a variety of different use-cases. Through our formative and user study, we identify some potential use cases.

Prior-visit auditing. Uncertainty about a space's accessibility often deterred participants from engaging in activities and made potential visits more daunting and stressful. Accessibility Scout is a practical solution to this problem, allowing users to preview potential accessibility concerns before their visits. Users could use Accessibility Scout to plan what assistive devices to take, decide whether to travel with a partner or caretaker, or guide deeper inquiries into specific accessibility risks.

Prior-visit location selection. Difficulty finding accessible dining or lodging often discouraged participants from trying new places. Accessibility Scout simplifies this process by allowing users to run scans across on publicly available images of environments to generate easily skimmable accessibility insights that can guide future trips and uncover new places to visit. Accessibility Scout's scalability can also enable future work conducting large-scale analysis of the accessibility of built environments across regions.

Sharing lived experiences. Participants were excited that Accessibility Scout could help them share parts of their lived experiences that are often hard to describe, building empathy and understanding among those around them. In doing so, participants believed they could better prepare the people around them to be more informed accessibility auditors. For other people experiencing major life changes like illness or injury or their loved ones, Accessibility Scout can reduce the uncertainty of adapting by providing new perspectives on what challenges in built environments they may encounter in the future.

Improving spaces for new demographics. While existing accessibility tools like ADA checklists [3] or RASSAR [74] try to capture as many people as possible in its static definitions of accessibility, Accessibility Scout allows building owners and businesses to identify key accessibility concerns for specific demographics of people by evaluating environments on a specified user model, enabling a more targeted approach for space design. For instance, Airbnb owners can use Accessibility Scout to rearrange furniture for the specific needs of their guest or government officials can use Accessibility Scout to evaluate federal housing for target groups.

Perspectives on personalization. Our user study found that personalized data was perceived as more useful (Figure 10 left). With only a brief one-hour personalization session, Accessibility Scout generated user models appreciably different from baseline (Figure 8 left). We also note that participants had varying preferences for the way accessibility concerns are described. Given this, we believe that further personalization is necessary to adapt not only the data, but its representation to the preferences of the user. These results support the findings of previous works in perceived accessibility [23, 45, 52, 62, 76], which state that the willingness to travel to somewhere is highly dictated by an individual's perception of that environment's accessibility. Personalization offers a new approach to accessibility assessment by capturing a small slice of how they "see" to better measure whether or not they would really want to go. Beyond usability, participants also stated that personalization made them feel more heard. As P2 states, *"That's so cool because the user right away feels like they have a voice and they're being heard like hey this is a concern for me. So thats super cool"* (P2). Thus, we believe that personalization can be an important tool in building adoption for future AI accessibility systems by validating user experiences. While we take one approach using structured JSON to personalize LLM systems, future works should explore other representations and methods for user modeling like vector databases, retrieval-augmented generation, and post-training.

Limitations of using only images. Accessibility Scout was designed around only environmental images to leverage the wealth of publicly available data from the internet. However, relying solely on images also limits Accessibility Scout to only capturing general heuristics in comparison to exact measurements or environment dynamics. Thus, the quality of our predictions is limited by when an image was taken (*e.g.*, a path with snow in the winter vs. in the summer), how reflective the image is of the actual experience of being there (*e.g.*, professionally shot AirBnB pictures vs. user generated content), and what is shown. While participants in our user study felt that generated accessibility scans were detailed and accurate enough to be perceived as useful, we believe that Accessibility Scout serves more as a general-purpose tool to alert users to potential concerns over a comprehensive accessibility auditing system like RASSAR [74]. Further research in HCI and AI is needed to quantify and evaluate the quality of accessibility scan given personalization to build better ground truths and guide the development of future AI systems. We envision four future areas of work: 1) How can data accuracy from LLMs be improved through improved post-training, computer vision tooling, and model selection? 2) How can we improve an LLM's ability to make conjectures about non-visual properties from visual cues? 3) Accessibility Scout can be easily extended to include multiple images and other textual data by inputting more data into the context window. What kind of accessibility information can be collected to improve concern predictions other readily available data sources (*e.g.*, user reviews, booking location, time of visit)? 4) What level of assessment detail is needed for users to evaluate the accessibility of environments?

Explainability in AI systems for accessibility. Given that AI technologies are still new and unexplored, many participants in our user study were wary of new AI technologies, especially as many existing accessibility technologies were not applicable to their own needs. By designing Accessibility Scout to mimic how users think about accessibility assessment using a task affordance perspective, users are more able to engage with and understand how each concern was generated. Findings from our user study reflected this, where participants were more receptive to incorrect concerns as they could trace back the reason the concern was generated, felt more engaged in the accessibility assessment process, and generated concerns that they would actually encounter when entering the environment. This design principle closely follows Miller *et al.*'s [55] call to action for new explainable AI systems to avoid the "inmates running the asylum" problem, when systems are designed around researcher needs over the intended user's, and integrate existing models of how people generate, select, present, and evaluate explanations and decisions in AI systems. Prior work has also has demonstrated this, showing that AI trust and transparency are directly related [7, 42]. We believe that future LLMs systems, especially those that support the diverse needs of groups like people with disabilities, should continue to be designed around explainable prediction pipelines, which can make them more approachable for new users. In doing so, users can be more seamlessly integrated into future pipelines through human-AI collaborations, increasing the user tolerance for errors and hallucinations which are currently unavoidable in modern LLMs.

8 Conclusion

In this paper, we introduce Accessibility Scout, an AI system that offers a new approach to generating personalized accessibility scans at scale. Accessibility Scout uses human-AI collaborations to allow users to easily and effectively update their personalization by validating generated assessments. Accessibility Scout can take in images of any environment cheaply and quickly, making it uniquely equipped to conduct personalized accessibility scans at a greater scale. Our technical evaluations demonstrate that Accessibility Scout can effectively capture a wide range of different accessibility features and adapt to the varying needs of different users. Furthermore, our user studies demonstrate that not only did users find Accessibility Scout useful, but the addition of personalization enabled by Accessibility Scout improves the overall usability of the data. Through our work, we demonstrate how AI technologies can be used to build scalable personalized accessibility solutions by applying this approach to accessibility auditing, introducing new ways we can build inclusive spaces and technologies alike.

9 Acknowledgements

This research was supported by the University of Washington's Center for Research and Education on Accessible Technology and Experiences (CREATE) and the UCLA Digital Technology and Solutions OpenAI Project Grant. We give special thanks to our study collaborators and participants for their participation and feedback.

References

[1] 2025. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

[2] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. https://doi.org/10.48550/arXiv.1906.02569 arXiv:1906.02569 [cs]

[3] ADA Checklist. 2024. ADA Checklist for Existing Facilities. https://www.adachecklist.org/doc/fullchecklist/ada-checklist.pdf Accessed: 2025-03-12.

[4] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.

[5] Na Min An, Eunki Kim, Wan Ju Kang, Sangryul Kim, Hyunjung Shim, and James Thorne. 2025. Can LVLMs and Automatic Metrics Capture Underlying Preferences of Blind and Low-Vision Individuals for Navigational Aid? *arXiv preprint arXiv:2502.14883* (2025).

[6] Geethanjali Anjanappa. 2022. Deep Learning on 3D Point Clouds for Safety-Related Asset Management in Buildings. https://essay.utwente.nl/91463/.

[7] Zahra Atf and Peter R. Lewis. 2025. Is Trust Correlated With Explainability in AI? A Meta-Analysis. *IEEE Transactions on Technology and Society* (2025), 1–8. https://doi.org/10.1109/tts.2025.3558448

[8] J. Balado, L. Díaz-Vilariño, P. Arias, and M. Soilán. 2017. Automatic Building Accessibility Diagnosis from Point Clouds. *Automation in Construction* 82 (Oct. 2017), 103–111. https://doi.org/10.1016/j.autcon.2017.06.026

[9] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1007–1014. https://doi.org/10.1145/3604915.3608857 arXiv:2305.00447 [cs]

[10] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined AI personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[11] Bessler Daniel, Porzel Robert, Pomarlan Mihai, Beetz Michael, Malaka Rainer, and Bateman John. 2020. A Formal Model of Affordances for Flexible Robotic Task Execution. In *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/FAIA200374

[12] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[13] Virginia Braun and Victoria Clarke. 2019. Reflecting on Reflexive Thematic Analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[14] Centers for Disease Control and Prevention. 2019. Disability Impacts All of Us. https://www.cdc.gov/ncbddd/disabilityandhealth/infographic-disability-impacts-all.html. Accessed: 2025-04-08.

[15] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. 2015. Mining Semantic Affordances of Visual Object Categories. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 4259–4267. https://doi.org/10.1109/cvpr.2015.7299054

[16] Changmao Chen, Yuren Cong, and Zhen Kan. 2024. WorldAfford: Affordance Grounding Based on Natural Language Instructions. https://arxiv.org/abs/2405.12461v1.

[17] Guangran Cheng, Chuheng Zhang, Wenzhe Cai, Li Zhao, Changyin Sun, and Jiang Bian. 2024. Empowering Large Language Models on Robotic Manipulation with Affordance Prompting. https://arxiv.org/abs/2404.11027v1.

[18] Hao-Yun Chi, Jingzhen Sha, and Yang Zhang. 2023. Bring Environments to People–A Case Study of Virtual Tours in Accessibility Assessment for People with Limited Mobility. In *20th International Web for All Conference*. 96–103.

[19] Hea Young Cho, Malcolm MacLachlan, Michael Clarke, and Hasheem Mannan. 2016. Accessible Home Environments for People with Functional Limitations: A

Systematic Review. *International Journal of Environmental Research and Public Health* 13, 8 (Aug. 2016), 826. https://doi.org/10.3390/ijerph13080826

[20] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. 2018. Learning to Act Properly: Predicting and Explaining Affordances from Images. *arXiv* (Dec. 2018).

[21] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A. Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, Robby Bryant, Ryan G. Gomes, Allen Jiang, Roy Lee, Yun Liu, Javier Perez, Jameson K. Rogers, Cathy Speed, Shyam Tailor, Megan Walker, Jeffrey Yu, Tim Althoff, Conor Heneghan, John Hernandez, Mark Malhotra, Leor Stern, Yossi Matias, Greg S. Corrado, Shwetak Patel, Shravya Shetty, Jiening Zhan, Shruthi Prabhakara, Daniel McDuff, and Cory Y. McLean. 2024. Towards a Personal Health Large Language Model. https://doi.org/10.48550/arXiv.2406.06474 arXiv:2406.06474 [cs]

[22] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.

[23] Angela Curl, John D. Nelson, and Jillian Anable. 2015. Same Question, Different Answer: A Comparison of GIS-based Journey Time Accessibility with Self-Reported Measures from the National Travel Survey in England. *Computers, Environment and Urban Systems* 49 (Jan. 2015), 86–97. https://doi.org/10.1016/j.compenvurbsys.2013.10.006

[24] Thi Hong Diep Dao and Jean-Claude Thill. 2018. Three-Dimensional Indoor Network Accessibility Auditing for Floor Plan Design. *Transactions in GIS* 22, 1 (2018), 288–310. https://doi.org/10.1111/tgis.12310

[25] Valerie Van der Linden, Hua Dong, and Ann Heylighen. 2016. From Accessibility to Experience: Opportunities for Inclusive Design in Architectural Practice. *NA* 28, 2 (Oct. 2016).

[26] Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating automatic feedback on ui mockups with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[27] Valerie Fletcher, Gabriela Bonome-Sims, Barbara Knecht, Elaine Ostroff, Jennifer Otitigbe, Maura Parente, and Joshua Safdie. 2015. The Challenge of Inclusive Design in the US Context. *Applied Ergonomics* 46 (Jan. 2015), 267–273. https://doi.org/10.1016/j.apergo.2013.03.006

[28] Qiang Fu, Hongbo Fu, Hai Yan, Bin Zhou, Xiaowu Chen, and Xueming Li. 2020. Human-Centric Metrics for Indoor Scene Assessment and Synthesis. *Graphical Models* 110 (July 2020), 101073. https://doi.org/10.1016/j.gmod.2020.101073

[29] Feifan Gao, Hanbei Cheng, Zhigang Li, and Le Yu. 2024. Revisiting the Impact of Public Spaces on the Mental Health of Rural Migrants in Wuhan: An Integrated Multi-Source Data Analysis. *International Journal of Health Geographics* 23, 1 (March 2024), 7. https://doi.org/10.1186/s12942-024-00365-8

[30] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[31] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[32] C. Harrison, P. Dall, P. M. Grant, M. Granat, T. Maver, and B. Conway. 2000. Development of a wheelchair virtual reality platform for use in evaluating wheelchair access. https://www.semanticscholar.org/paper/Development-of-a-wheelchair-virtual-reality-for-use-Harrison-Dall/2204abde6e176ff746ffefa50a3b8f83696d5671

[33] Beverly P. Horowitz, Almonte , Tiffany, and Andrea and Vasil. 2016. Use of the Home Safety Self-Assessment Tool (HSSAT) within Community Health Education to Improve Home Safety. *Occupational Therapy In Health Care* 30, 4 (Oct. 2016), 356–372. https://doi.org/10.1080/07380577.2016.1191695

[34] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (March 2025), 1–55. https://doi.org/10.1145/3703155 arXiv:2311.05232 [cs]

[35] William Huang, Sam Ghahremani, Siyou Pei, and Yang Zhang. 2024. WheelPose: Data Synthesis Techniques to Improve Pose Estimation Performance on Wheelchair Users. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.

[36] Koichi Ito and Filip Biljecki. 2021. Assessing Bikeability with Street View Imagery and Computer Vision. *Transportation Research Part C: Emerging Technologies* 132 (Nov. 2021), 103371. https://doi.org/10.1016/j.trc.2021.103371

[37] Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 796–806. https://doi.org/10.1145/3626772.3657815

[38] Nikolaos Kaklanis, Panagiotis Moschonas, Konstantinos Moustakas, and Dimitrios Tzovaras. 2013. Virtual User Models for the Elderly and Disabled for Automatic Simulated Accessibility and Ergonomy Evaluation of Designs. *Universal Access in the Information Society* 12 (Nov. 2013). https://doi.org/10.1007/s10209-012-0281-0

[39] Efthimis Kapsalis, Nils Jaeger, and Jonathan Hale. 2024. Disabled-by-Design: Effects of Inaccessible Urban Public Spaces on Users of Mobility Assistive Devices – a Systematic Review. *Disability and Rehabilitation: Assistive Technology* 19, 3 (April 2024), 604–622. https://doi.org/10.1080/17483107.2022.2111723

[40] Piyawan Kasemsuppakorn, Hassan A. Karimi, Dan Ding, and Manoela A. Ojeda. 2015. Understanding Route Choices for Wheelchair Navigation. *Disability and Rehabilitation. Assistive Technology* 10, 3 (May 2015), 198–210. https://doi.org/10.3109/17483107.2014.898160

[41] William Kirk, August Lösch, and Isaiah Berlin. 1963. Problems of Geography. *Geography* 48, 4 (1963), 357–371. jstor:40565711

[42] Bran Knowles, John T. Richards, and Frens Kroeger. 2022. The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency. https://doi.org/10.48550/arXiv.2208.00681 arXiv:2208.00681 [cs]

[43] Werner Kurschl, Mirjam Augstein, Thomas Burger, and Claudia Pointner. 2014. User Modeling for People with Special Needs. *International Journal of Pervasive Computing and Communications* 10, 3 (Jan. 2014), 313–336. https://doi.org/10.1108/IJPCC-07-2014-0040

[44] Dan Lämkull, Lars Hanson, and Roland Ortengren. 2007. The Influence of Virtual Human Model Appearance on Visual Ergonomics Posture Evaluation. *Applied Ergonomics* 38, 6 (Nov. 2007), 713–722. https://doi.org/10.1016/j.apergo.2006.12.007

[45] Katrin Lättman, Lars E. Olsson, and Margareta Friman. 2018. A New Approach to Accessibility – Examining Perceived Accessibility in Contrast to Objectively Measured Accessibility in Daily Travel. *Research in Transportation Economics* 69 (Sept. 2018), 501–511. https://doi.org/10.1016/j.retrec.2018.06.002

[46] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. 2023. Semantic-SAM: Segment and Recognize Anything at Any Granularity. https://doi.org/10.48550/arXiv.2307.04767 arXiv:2307.04767

[47] Hong Li, Na Ta, Bailang Yu, and Jiayu Wu. 2023. Are the Accessibility and Facility Environment of Parks Associated with Mental Health? A Comparative Analysis Based on Residential Areas and Workplaces. *Landscape and Urban Planning* 237 (Sept. 2023), 104807. https://doi.org/10.1016/j.landurbplan.2023.104807

[48] Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, and Feifei Feng. 2024. Improving Vision-Language-Action Models via Chain-of-Affordance. https://doi.org/10.48550/arXiv.2412.20451 arXiv:2412.20451 [cs]

[49] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2023. ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. https://doi.org/10.48550/arXiv.2305.06566 arXiv:2305.06566 [cs]

[50] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. https://doi.org/10.48550/arXiv.2307.15780 arXiv:2307.15780 [cs]

[51] Lena Mamykina, Daniel A. Epstein, Predrag Klasnja, Donna Sprujt-Metz, Jochen Meyer, Mary Czerwinski, Tim Althoff, Eun Kyoung Choe, Munmun De Choudhury, and Brian Lim. 2022. Grand challenges for personal informatics and AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.

[52] Gavin R. McCormack, Ester Cerin, Eva Leslie, Lorinne Du Toit, and Neville Owen. 2008. Objective Versus Perceived Walking Distances to Destinations: Correspondence and Predictive Validity. *Environment and Behavior* 40, 3 (May 2008), 401–425. https://doi.org/10.1177/0013916507300560

[53] Mike A. Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y. McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. 2024. Transforming Wearable Data into Health Insights Using Large Language Model Agents. https://doi.org/10.48550/arXiv.2406.06464 arXiv:2406.06464 [cs]

[54] Sara S. Metcalf, Mary E. Northridge, Michael J. Widener, Bibhas Chakraborty, Stephen E. Marshall, and Ira B. Lamster. 2013. Modeling Social Dimensions of Oral Health Among Older Adults in Urban Environments. *Health education & behavior : the official publication of the Society for Public Health Education* 40, 1 0 (Oct. 2013), 63S–73S. https://doi.org/10.1177/1090198113493781

[55] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. https://doi.org/10.48550/arXiv.1712.00547 arXiv:1712.00547 [cs]

[56] Huaqing Min, Chang'an Yi, Ronghua Luo, Jinhui Zhu, and Sheng Bi. 2016. Affordance Research in Developmental Robotics: A Survey. *IEEE Transactions on Cognitive and Developmental Systems* 8, 4 (Dec. 2016), 237–255. https://doi.org/10.1109/TCDS.2016.2614992

[57] J. M. Morris, P. L. Dumble, and M. R. Wigan. 1979. Accessibility Indicators for Transport Planning. *Transportation Research Part A: General* 13, 2 (April 1979), 91–109. https://doi.org/10.1016/0191-2607(79)90012-8

[58] Abdelhak Moussaoui, Alain Pruski, and Choubeila Maaoui. 2012. Virtual Reality for Accessibility Assessment of a Built Environment for a Wheelchair User. *Technology and Disability* 24, 2 (May 2012), 129–137. https://doi.org/10.3233/

TAD-2012-0341

[59] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109* (2024).

[60] Siyou Pei, Alexander Chen, Chen Chen, Franklin Mingzhe Li, Megan Fozzard, Hao-Yun Chi, Nadir Weibel, Patrick Carrington, and Yang Zhang. 2023. Embodied Exploration: Facilitating Remote Accessibility Assessment for Wheelchair Users with Virtual Reality. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3597638.3608410

[61] Emiliano Pérez, Alejandro Espacio, Santiago Salamanca, and Pilar Merchán. 2022. WUAD (Wheelchair User Assisted Design): A VR-Based Strategy to Make Buildings More Accessible. *Applied Sciences* 12, 17 (Jan. 2022), 8486. https://doi.org/10.3390/app12178486

[62] Felix Johan Pot, Bert van Wee, and Taede Tillema. 2021. Perceived Accessibility: What It Is and Why It Differs from Calculated Accessibility Measures Based on Spatial Data. *Journal of Transport Geography* 94 (June 2021), 103090. https://doi.org/10.1016/j.jtrangeo.2021.103090

[63] Wolfgang F. E. Preiser, Jacqueline Vischer, and Edward White. 2015. *Design Intervention (Routledge Revivals): Toward a More Humane Architecture.* Routledge.

[64] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. 2024. AffordanceLLM: Grounding Affordance from Vision Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Seattle, WA, USA, 7587–7597. https://doi.org/10.1109/cvprw63382.2024.00754

[65] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[66] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, and Jon Froehlich. 2019. Project Sidewalk: A Web-based Crowdsourcing Tool for Collecting Sidewalk Accessibility Data At Scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300292

[67] Mark C. Schall Jr., Nathan B. Fethke, and Victoria Roemig. 2018. Digital Human Modeling in the Occupational Safety and Health Process: An Application in Manufacturing. *IISE Transactions on Occupational Ergonomics and Human Factors* 6, 2 (April 2018), 64–75. https://doi.org/10.1080/24725838.2018.1491430

[68] Tom Seekins, Meg A. Traci, and Emily C. Hicks. 2022. Exploring Environmental Measures in Disability: Using Google Earth and Street View to Conduct Remote Assessments of Access and Participation in Urban and Rural Communities. *Frontiers in Rehabilitation Sciences* 3 (Aug. 2022), 879193. https://doi.org/10.3389/fresc.2022.879193

[69] Andrés Serna and Beatriz Marcotegui. 2013. Urban Accessibility Diagnosis from Mobile Laser Scanning Data. *ISPRS Journal of Photogrammetry and Remote Sensing* 84 (Oct. 2013), 23–32. https://doi.org/10.1016/j.isprsjprs.2013.07.001

[70] Ather Sharif, Paari Gopal, Michael Saugstad, Shiven Bhatt, Raymond Fok, Galen Weld, Kavi Asher Mankoff Dey, and Jon E. Froehlich. 2021. Experimental Crowd+AI Approaches to Track Accessibility Features in Sidewalk Intersections Over Time. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3441852.3476549

[71] Hyun Oh Song, Mario Fritz, Daniel Goehring, and Trevor Darrell. 2016. Learning to Detect Visual Grasp Affordance. *IEEE Transactions on Automation Science and Engineering* 13, 2 (April 2016), 798–809. https://doi.org/10.1109/tase.2015.2396014

[72] Xia Su, Daniel Campos Zamora, and Jon E Froehlich. 2024. RAIS: Towards A Robotic Mapping and Assessment Tool for Indoor Accessibility Using Commodity Hardware. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–5.

[73] Xia Su, Ruiqi Chen, Weiye Zhang, Jingwei Ma, and Jon E Froehlich. 2024. A Demo of DIAM: Drone-based Indoor Accessibility Mapping. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–3.

[74] Xia Su, Han Zhang, Kaiming Cheng, Jaewook Lee, Qiaochu Liu, Wyatt Olson, and Jon E Froehlich. 2024. RASSAR: Room Accessibility and Safety Scanning in Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[75] Jacqueline Vaughn Switzer. 2003. *Disabled rights: American disability policy and the fight for equality.* Georgetown University Press.

[76] Anna-Lena van der Vlugt, Angela Curl, and Dirk Wittowsky. 2019. What about the people? Developing measures of perceived accessibility from case studies in Germany and the UK. *Applied Mobilities* 4, 2 (May 2019), 142–162. https://doi.org/10.1080/23800127.2019.1573450 Publisher: Routledge _eprint: https://doi.org/10.1080/23800127.2019.1573450.

[77] Galen Weld, Esther Jang, Anthony Li, Aileen Zeng, Kurtis Heimerl, and Jon E. Froehlich. 2019. Deep Learning for Automatically Detecting Sidewalk Accessibility Problems Using Streetscape Imagery. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 196–209. https://doi.org/10.1145/3308561.3353798

[78] Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. *arXiv preprint arXiv:2402.09269* (2024).

[79] Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yuping Jin, Shi Chen, and Lingyun Sun. 2024. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[80] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv:2310.11441 [cs]

[81] Bangpeng Yao, Jiayuan Ma, and Li Fei-Fei. 2013. Discovering Object Functionality. In *2013 IEEE International Conference on Computer Vision*. IEEE, Sydney, Australia, 2512–2519. https://doi.org/10.1109/iccv.2013.312

[82] Matteo Zallio and P. John Clarkson. 2021. Inclusion, Diversity, Equity and Accessibility in the Built Environment: A Study of Architectural Design Practice. *Building and Environment* 206 (Dec. 2021), 108352. https://doi.org/10.1016/j.buildenv.2021.108352

[83] Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. LLM-based Medical Assistant Personalization with Short- and Long-Term Memory Coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2386–2398. https://doi.org/10.18653/v1/2024.naacl-long.132

[84] Qiuyi Zhang, Mary E. Northridge, Zhu Jin, and Sara S. Metcalf. 2018. Modeling Accessibility of Screening and Treatment Facilities for Older Adults Using Transportation Networks. *Applied geography (Sevenoaks, England)* 93 (April 2018), 64–75. https://doi.org/10.1016/j.apgeog.2018.02.013

[85] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024. Personalization of Large Language Models: A Survey. https://doi.org/10.48550/arXiv.2411.00027 arXiv:2411.00027 [cs]

[86] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2025. Personalization of Large Language Models: A Survey. https://doi.org/10.48550/arXiv.2411.00027 arXiv:2411.00027 [cs]

A Formative Study Demographics

Table A.1: Demographic information of six participants (U1-U6) in the formative study.

					No. Preferred			
ID	Age	Gender	Diagnosed Disability	Self-Described Motor Capability	Total	Personalized	Equal	Generic
U1	41	M	Paraplegic	Able to use hands fully, paralyzed from the waist down	4	3	1	0
U2	29	F	Leg amputation	Limited to moving around certain terrains	5	2	1	2
U3	53	M	Spinal cord injury - T11/T12, paraplegic	Over long distances I use my wheelchair, for short distances I use crutches	4	3	1	0
U4	28	F	Multiple sclerosis	Numbness on my arms and feet, a lot of fatigue	4	2	2	0
U5	47	M	C6 incomplete quadriplegic	No movement from the chest down, C6 and below	4	2	1	1
U6	50	M	Spinal cord injury - C5/C6	Paralysis below mid chest, limited hand movement, limited wrist flexion, no triceps	2	0	2	0

B Prompts

```
##INSTRUCTIONS##
You are tasked with identifying the potential tasks a user might perform in a given
↪  space. You will be given a set of images and a brief textual description of the
↪  environment and what the user intends to do. Identify all the potential tasks that
↪  might be performed within the environment depicted in the pictures given the
↪  provided description and items in the environment. Be as concise as possible.
↪  Describe only the most relevant tasks. Do not add any tasks that would be
↪  extraneous. Respond in JSON with these keys and values: "name": string, name of
↪  the task, "desc": string, brief description of what the task involves.

##EXAMPLES##
Input: an image of a bathroom
Output: [
    {
        "name": "Using the Toilet",
        "desc": "Using the toilet for personal needs"
    },
    {
        "name": "Washing Up",
        "desc": "Washing your face and body and freshening up in the morning"
    },
    {
        "name": "Taking care of Oral Hygiene",
        "desc": "Brushing teeth and rinsing your mouth"
    }
]

Input: An image of a restaurant I am going on a date at
Output: [
    {
        "name": "Dining",
        "desc": "Eating comfortably at the restaurant"
    },
    {
        "name": "Reading the Menu",
        "desc": "Checking the menu to know what to order"
    },
    {
        "name": "Chatting",
        "desc": "Talking with your date
    }
]
```

Figure B.1: Prompt used to identify subtasks a user might do in an environment.

```
##INSTRUCTIONS##
You are tasked with identifying all the possible locations a user might perform a task
↪  in. You will be given an image and environment description and a task in JSON form
↪  with a name field and a brief description in the desc field. Identify potential
↪  locations in the image that the user may need to interact with to perform the task.
↪  Be as concise as possible, describing only the most important locations. Respond
↪  in JSON with these keys and values: "name": string, name of the location, "reason":
↪  string, why the user will interact with this location, "primitives": list[string],
↪  a list of all primitive motions or actions the user may need to do to perform the
↪  task at this location. This should be as exhaustive as possible while only listing
↪  general motions. These should all be motions or physical actions. For example,
↪  primitives could include "reaching arm up" or "sitting down". These should be as
↪  general of motions as possible while describing what the user might perform.

##EXAMPLES##
Input: A picture of a bathroom.
{
    "name": "Using the Toilet",
    "desc": "Using the toilet for personal needs"
}

Output:
[
    {
        "name": "toilet",
        "reason": "Conduct personal needs"
        "primitives": [
            "sit down",
            "stand up",
            "bend over"
        ],
        "name": "Sink",
        "primitives": [
            reach with arm,
            grasp,
        ]
    }
]

Input: A picture of a restaurant
{
    "location": "Dining",
    "desc": "Eating comfortably at the restaurant"
}

Output:
[
    {
        "location": "table",
        "reason": "Food will be served at the table"
        "primitives": [
            "sit down",
            "stand up",
            "grasp",
            "read in dark"
        ]
    }
]
```

Figure B.2: Prompt used to identify locations user might perform a subtask and their primitive motions.

```
##INSTRUCTIONS##
You are a accessibility practioner who is well versed and knowledgeable about mobility
↪  limitations and their potential implications. You are tasked with assessing the
↪  accessibility of different tasks within an environment. The user will give you a
↪  description of their physical abilities and conditions in the form of a high level
↪  description and a JSON with these keys and values in order: "name": string, short
↪  name for the basic movement; "desc": string, one sentence description why this
↪  movement is affected, "frequent": bool, true if this movement is common in
↪  everyday life, "affected_part": a string of the body part this may affect from
↪  (arms, legs, feet back, chest, hands, eyes, ears, brain, user preference). Do not
↪  answer with any hypotheticals. Assess the accessibility of performing the
↪  specific task or action in this environment. Only give accessibility concerns for
↪  parts of the environment that would affect the user from performing the task. Do
↪  not give any concerns for anything this is not relevant to completing the task. A
↪  concern can also be the lack of something like grab bars or handle bars. A concern
↪  can also be the size or shape of the space. You can respond with empty JSONs if
↪  there are no concerns. Contextualize all your answers to what the user can and
↪  can't do. Always justify a concern by one of the given user capabilities. Concerns
↪  should focus directly on the environment. Only label concerns you are certain
↪  would be an issue. Do not use words like "may", "if", or "potentially". You will
↪  then be given an image with number annotations to reference different parts of the
↪  environment and a description of the environment this image is of. Respond in JSON
↪  with these keys and values in order: "name": string, name which is descriptive of
↪  the exact environment concern, "desc": string, brief description of why this
↪  concern would affect the user with no mention of any annotated numbers,
↪  "locations": list[int], the number on the image that is annotated on top of the
↪  concern. Answer only the number mark closest to the concern. Ignore the presence
↪  of people and only focus on aspects of the physical environment.

##EXAMPLES##
Inputed User Model:
[
    {
        "name": "Walking",
        "desc": "The user cannot perform this movement due to reliance on a wheelchair
        ↪  for mobility.",
        "frequent": true,
        "affected_part": "legs"
    },
    {
        "name": "Running",
        "desc": "The user is unable to perform running due to limitations requiring a
        ↪  wheelchair.",
        "frequent": true,
        "affected_part": "legs"
    },
    {
        "name": "Stair Climbing",
        "desc": "The user cannot climb stairs as it requires leg strength and mobility
        ↪  that are impaired.",
        "frequent": true,
        "affected_part": "legs"
    },
    {
        "name": "Standing",
        "desc": "The user is unable to stand independently due to limitations in leg
        ↪  support and balance.",
        "frequent": true,
        "affected_part": "legs"
    }
]

Input: A picture of a bathroom
Output:
[
    {
        "name": "Slippery Floors",
        "desc": "The marble on the floors can be slippery making it hard to push a
        ↪  wheelchair",
        "locations": [
            3,
            4
        ]
    },
    {
        "name": "High Bathtub Walls",
        "desc": "The user can not get into the bathtub due to wheelchair usage",
        "locations: [
            8
        ]
    },
    {
        "name": "High Mirror",
        "desc": "User is too low to see mirror when in wheelchair",
        "locations": [
            15
        ]
    },
    {
        "name": "Out of Reach Outlet",
        "desc": "Outlet is too far to reach from wheelchair",
        "locations": [
            19
        ]
    }
]
```

Figure B.3: Prompt used to identify accessibility concerns in an image.

C Thematic Coding

Table C.2: All generated themes and codes from user study.

Theme	Code	Count
User Practices	Existing Practices For Accessibility Evaluations	4
	People Evaluate Accessibility For Their Entire Party	1
General Concerns About Environment	Children Compatible Design	1
	Concerns Captured by ADA	9
	Concerns Due to Existence of Other People	4
	Accessibility Information Availability	1
	ADA Is Not Fully Enforced or Maintained	2
System Usefulness	Evaluations Have Correlation With Planned Activity	6
	Participants Expect Inaccessibilities	4
	Context Influences Accessibility Concerns	10
	System Identifies Hard to Notice Concerns	2
	Can/Cannot Infer Non-Visual Properties from Visual	6
	Knowing Accessibility Is More Useful When Choosing a Locations	1
	Data Availability Affects Usefulness of System	6
	Commonly Accepted Concerns Are Not Useful	2
	Build Understanding if Disabilities	6
System Features	Able to Accept New Concerns	1
	Structured Presentation of Information	3
	Training Process Is Positively Perceived	7
	Image and Image Highlighting Are Useful	3
	Usefulness of Concern Reasoning	3
	System Is Fast and Responsive	2
Improvements Needed	Overfitting	1
	Delay	1
	Hallucinations	2
	Highlighting Is Inaccurate	3
	Duplicated Concerns	1
	User Needs Are Dynamic	4
	Detail Level of Concern Descriptions	8
	Accuracy And Detail of Concern Reasoning	13
	Control of What Concerns Are Shown	2
	Accessibility Accomodation Recommendations	1
	Clickable Highlights	1
	Directly Query AI About Places	1
	Integrate Other Human Feedback	2
	Use Reference Measurements in Concern Descriptions	1
AI Usage	AI Usage is Appropriate for Accessibility Scans	9
	Concerns on AI Accuracy	4
	Concerns on Data Security in AI	6
	AI Capabilities of Capturing Perception	4
	Fear Of AI Influencing Perception	3
	AI Mimicking User Perception Is Useful in Accessibility Scans	3
	AI Mimimcking Perception Allows Word View to be Shared	1
	Importance Of Mobility Modeling Accuracy in AI	5
	Lack Of Trust in AI Prescriptions Of Mobility	4
Personalization	Personalization Builds Trust in AI	2
	Everyone Is Unique	4
	Concerns Should Mention User Capabilities	9
	Personalization Make People Feel Heard	2
	Personalization Reminds People of Their Disabilities and Makes Them Uncomfortable	3
	Mixed Feelings On Usefulness of Personalization	8