tvst

Low Vision Rehabilitation

Using ChatGPT To Generate Walking Directions—Could It Potentially Help Visually Impaired With Micronavigation?

Junxian He[1,2], Shrinivas Pundlik[1], and Gang Luo[1]

[1] Schepens Eye Research Institute of Mass Eye & Ear, Harvard Medical School, Boston MA, USA
[2] Chongqing University, Chongqing, China

Correspondence: Gang Luo, Schepens Eye Research Institute of Mass Eye & Ear, Harvard Medical School, Boston, MA 02114, USA. e-mail: gang.luo@schepens.harvard.edu

Received: September 6, 2024
Accepted: March 14, 2025
Published: May 28, 2025

Keywords: chatbot; ChatGPT; multi-modal LLM; navigation; micronavigation; visual impairment

Citation: He J, Pundlik S, Luo G. Using ChatGPT to generate walking directions—could it potentially help visually impaired with micronavigation? Transl Vis Sci Technol. 2025;14(5):28, https://doi.org/10.1167/tvst.14.5.28

Purpose: Micronavigation poses challenges for blind and visually impaired individuals (BVI). We investigated the potential of using ChatGPT4o to provide navigation directions.

Methods: We created a dataset of real-world micronavigation scenarios consisting of 113 scene images and corresponding human-generated scene descriptions. A total of 407 human-generated wayfinding queries and their expected responses were used for evaluation (245 answerable and 162 not answerable based on the scenes). "I do not know" as a response was expected for the negative cases. Sensitivity (SEN) and specificity (SPE) of navigation directions output by ChatGPT4o was evaluated under different input conditions.

Results: The default ChatGPT4o, with scene images as inputs, resulted in SEN and SPE values of 64% and 75.9%, respectively. Instructing via prompts on how to respond to unanswerable questions did not affect SEN, but increased SPE by 14 %. SEN and SPE both decreased by about 13% and 7%, respectively, when ChatGPT4o first generated text description from the input images and then answered the queries based the generated descriptions. When human-written scene descriptions were provided as input, SEN and SPE improved by about 17% and 16%, respectively.

Conclusions: Some spatial information needed for micronavigation was diminished in generative scene descriptions. This problem may be addressed by one-step vision language models. General-purpose ChatGPT4o with fine-tuned prompts was still inferior to human in scene understanding. A multimodal large language model custom-trained on navigation scenarios might help improve the performance.

Translational Relevance: AI chat-bots may have the potential to provide navigation assistance to BVI to a certain degree.

Introduction

Blind and visually impaired individuals (BVI) often face challenges related to orientation and mobility in their daily-life tasks. Navigation, or wayfinding, is one of the key components of mobility. Assistive devices and technologies, including mobile apps, are intended to help the BVI with a wide-variety of daily-life tasks, including navigation.[1] Navigation can be further classified as macro- or micronavigation. When one performs a navigation task of getting from point A to point B in its entirety, macronavigation refers to the high-level aspects of path/route planning and following-up, generally over larger distances (say over many miles), and often facilitated by geo-localization/mapping technologies.[2,3] Micronavigation, on the other hand, refers to navigating within a close range (sometimes within a few meters) during the journey—precisely to a particular location, for instance, finding store entrances, train station exits, and so on. Planning and execution of micronavigation tasks is often ignored or taken for granted in the general context of navigation by sighted humans. Although a macronavigation tool can guide BVI individuals to the vicinity of a building, yet getting into the building or getting to the elevator entrance, can be challenging micronavigation tasks that are also essential aspects

translational vision science & technology

of the entire trip. However, this is exactly the kind of the task that is largely unresolved for BVI individuals who want to travel independently,[4,5] because there is no general-purpose micronavigation tool or aid and they frequently have to resort to asking others for help.

Although there are a large number of devices and smartphone apps for navigation assistance, overwhelming number of them are for macronavigation.[2,3] Mapping and geolocation based apps are not particularly adequate for micronavigation, not only because of the inherent errors in geolocalization and mapping,[6] but also because of an overall lack of mapping or insufficient mapping, as it is not feasible to accurately map all the locations. Places away from streets or inside buildings are often not mapped. Some vision aids and assistance apps make use of computer vision algorithms to perform object detection[1] or provide micronavigation assistance in certain specific scenarios—such as public transit specific information.[7,8] Because of the sheer variety of objects one could encounter in the real world, the development of custom tools for many kinds of specific objects still has a long way to go. Moreover, using customized solutions for different scenarios, for instance, navigating to bus stops, has limited operational feasibility from the perspective of BVI users. Apps and services such as Aira, which provide live, remote personalized assistance to BVI individuals could be helpful in navigation.[9–12] However, cost and feasibility of human assistance means that its utilization tends to be limited. A 24 × 7 virtual assistant could alleviate many of these micronavigation related challenges for BVI individuals. Such a virtual assistant for BVI travelers is not out of the realm of possibility, given the recent advances in artificial intelligence models for computer vision, large language models (LLMs), and visual-language models (VLMs).[13,14]

VLMs are designed for tasks that require some combination of computer vision and natural-language inputs/outputs such as image captioning or visual question answering, among others.[13] VLMs have been used for navigation, especially for robot navigation, where the idea is that the robot will take natural language instructions, extract salient information from the text (such as landmarks and their inter-relationship), and then perform the navigation tasks based on visual detection of the said landmarks in previously unseen environments.[15–25] Although the vision-language navigation approaches are focused understanding image- and language-based cues for navigation by autonomous agents, our goal in this study was to evaluate whether general purpose foundational models such as ChatGPT 4o have the ability to prove direct and precise responses to micronavigation-related queries of BVI users.

The task of providing navigation information by an AI agent involved two components: (1) scene understanding, where navigation-relevant information needs to be extracted from input images of the scene; and (2) generation of specific instructions/directions for the human users based on natural language interaction.

Abilities of AI models to interact with humans are now evident, with the advent of AI agents like ChatGPT in our day-to-day lives. However, scene understanding functionalities of these AI agents is relatively new, for example ChatGPT 4o with image inputs[26] or a large language-and-vision assistant.[19] Their ability to understand navigation-specific scenarios and provide clear instructions to humans is not fully evident as yet.

Be My AI function in the Be My Eyes app was introduced as a service for visually impaired individuals in scene understanding tasks, where a picture of the scene is interpreted by OpenAI's model to generate scene description. Other similar efforts such as Aira's initiative to build AI datasets to assist BVI users are also ongoing. Because these are relatively new services, rigorous studies and scientific reports about their effectiveness are scarce. Also, because these are designed to be a general purpose scene understanding agents/services, their effectiveness in navigation-specific scenarios is unknown.

In our evaluation study, we created a dataset of navigation scenarios by capturing a variety of real-world scene images. Scene description for navigation purposes, based on the captured pictures, was written by humans. Multiple positive (answerable) and negative (non-answerable) query-response pairs specific to each scenario (human-generated) were created. From there on, we tried various combinations of different input data formats and prompt-based training approaches to ChatGPT 4o with the goal of eliciting actionable responses from the agent. We compared the accuracy of the navigation directions to our queries provided by the AI-agents with the human generated responses. Our goal in this study was to understand feasibility and limitations of using AI agents like ChatGPT 4o in micronavigation scenarios. It did not involve BVI users, but this early-stage study can guide future developments in this field, including whether and how human subject studies may need to be conducted.

Methods

Dataset Generation

The dataset consisted of 113 images of real-world locations, both indoor and outdoor, represent-

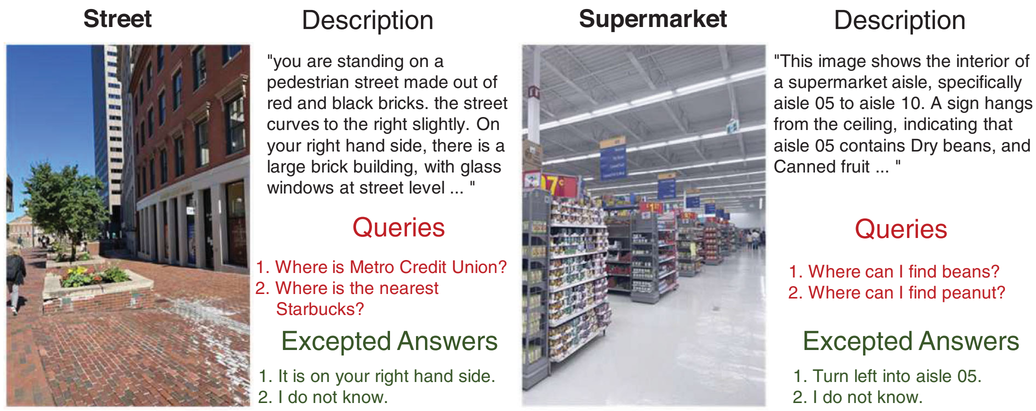


Figure 1. The dataset of navigation scenarios consisted of 113 indoor and outdoor images along with their human-generated text description of the scene, navigation-related queries based on the scene (positive and negative), and the expected answers to the queries. The human scene descriptions were written without knowing what the destination would be. Queries were composed based on images alone without seeing the description.

Table 1. Details About Types of Navigation Scenarios and Task Objectives

Type	Categories	Count
Scenarios (Total 113)	Street	45
	Indoor	42
	Transit stations	13
	Plaza	13
Destination/Target/	Specific Object	68
Task Objective	Direction/orientation	119
(Total 407)	Entrance	220

ing typical navigation scenarios, along with human generated textual descriptions of the scene depicted in these images (Fig. 1). Scenarios were broadly categorized into 4 categories: outdoor, indoor (mall, office, supermarket, etc.), transit stations (bus and subway), and plaza. Destinations were broadly categorized into three types: specific objects (door, chair, trash bin, etc.), directions to places that were typically indicated by signage and entrances to stores and buildings. Table 1 shows the details about scenarios and destination counts. The complexity of navigation tasks ranged from simple identification of landmarks to more complex scenarios involving turns. Obstacle avoidance was not specifically evaluated because we assume it can be addressed by the users' typical habitual mobility aids (such as long cane). One of the underlying assumptions was that the potential BVI users are independently mobile.

The images were captured during daytime with mobile phones by normally sighted individuals from the perspective of a potential user who would like to query navigation-related information. Short descriptions of the captured picture, typically a few lines, were written by members of the study team. The description included the general characteristics of the scene as well as specific inter-relationships of the various objects present in the scene. Emphasis was provided to make the description informative and relevant to navigation-related queries that may potentially arise. However, the description was agnostic to navigation destination(s) involved in the scenarios. Additionally, multiple navigation-related queries were written by the study team for each scene image (based on the scene image alone without the knowledge of its corresponding text description), some were answerable while some were not answerable based on the information in the captured scene. Answerable queries were those related to the information present in the scene image and the corresponding text description. Thus, a person looking at the picture of the scene and/or text description of the scene could reasonably and accurately answer the query and provide specific navigation related guidance. On the other hand, unanswerable queries were those for which information was not present and the expected answer was "I do not know." Ground truth responses (expected responses written by humans) to all the queries were evaluated by the study staff for clarity and correctness. Moreover, responses were crafted such that they did not rely entirely on visual elements in the scene. From 113 images and their associated text descriptions, 407 total queries and their template responses were created.

Micronavigation Prompts

Using OpenAI's API (application programming interface), we created a navigation chat bot based on ChatGPT 4o model. A Python code was created to automatically submit queries and receive responses. We

translational vision science & technology

Table 2. Prompts Provided to the ChatGPT 4o MicroNavigation Assistants

Prompt Label	Prompt
Prompt 1	N/A
Prompt 2	You are a navigator that should give correct directions to help a person go to where they need to go
Prompt 3	You are a navigator specifically made to help people with vision disabilities or blind people. Your task is to suggest walking directions to when responding to a navigation question in the user message. Before you answer, please think step by step. Your answer should be correct. If the information provided is not enough to form a full set of directions to the destination, do not output any directions

experimented with many prompts and finally chose 3 different versions of prompts (Table 2) to instruct the chat bot to work as micronavigation assistant. In choosing the prompts, we utilized a universal prompt generation framework suggested by OpenAI,[27] which includes the following: (1) clearly defining the objective, (2) providing background information, and (3) using specific instructions. We intended to follow a natural progression in our prompt selection, ranging from no prompt (as indicated by Prompt 1 in Table 2) to progressively clearer and more refined prompts (Prompts 2 and 3 in Table 2). We implemented strategies 1 and 2 in Prompt 2, whereas we incorporated all three strategies in Prompt 3. These prompts can be provided via the "instructions" variable in the python code or via the "Instructions" text box in the OpenAI assistant UI (see Appendix I). Two members of our team separately tried seven different prompts on a subset of the dataset, and the best prompt based on SEN and the best prompt based on SPE were selected for evaluation of the AI agents. A third prompt (default) was chosen to evaluate the effect of supplying a prompt.

Each of the three prompt instruction sets were used under three different input formats for generating navigation responses: (i) scene image only—where ChatGPT 4o generated navigation directions to the queries directly based on the scene images as input; (ii) human generated text descriptions of the scene images—where instead of the scene images, human generated text descriptions of the scenes were provided to ChatGPT to obtain navigation directions to the queries; and (iii) a two-step procedure where scene descriptions in text were first generated by ChatGPT 4o based on the input images, and the responses to the queries were obtained based on the generated text descriptions. Thus there were nine different combinations of input formats and prompt instructions, such that each could be considered as a different version of virtual navigation assistants.

The scene descriptions generated by ChatGPT 4o were very detailed as no specific query is provided. Further inference will be needed when answering specific wayfinding questions. Taking the Street picture in Figure 1 as an example, the description from ChatGPT was as follows:

> This image shows a vibrant urban scene in what appears to be a commercial or historical district. Here's a detailed description for wayfinding purposes. General Scene Description: The image shows a red-brick pedestrian walkway, flanked by buildings and decorative planters. The setting appears to be an open plaza or pedestrian area in a city, with clear skies and bright sunlight. To the right is a multi-story red-brick building housing businesses, including a clearly visible Metro Credit Union branch with signage. Walkway: The ground is paved with red bricks, creating a uniform, visually appealing path. White markings (possibly water or paint) are present on parts of the brick, but they don't obstruct movement. Planters: Large rectangular brick planters are positioned along the walkway. Each planter has vibrant greenery and flowers, including leafy plants and colorful blossoms, adding a welcoming touch to the area. Buildings: On the right side: A red-brick building has large windows, arched frames, and decorative trim. A Metro Credit Union branch is visible with signage and a 24-hour ATM. In the background: A tall, modern skyscraper contrasts with the historical feel of the red-brick building. A smaller historical-looking building with a triangular pediment can be seen at the end of the walkway. Landmarks: The Metro Credit Union branch is a notable point for orientation. The tall modern skyscraper provides a recognizable skyline landmark. The smaller historic building in the distance may serve as a destination or point of interest. People: Pedestrians are walking along the path, suggesting it's a public space with regular foot traffic. Wayfinding Details: If you're navigating this space: Use the Metro Credit Union building as a reference point. The credit union entrance is accessible from the walkway. The planters and brick pathway guide you forward toward the smaller historic building in the background. The open area and prominent buildings make it easy to orient yourself based on visual landmarks. Environmental Details: The sunny weather and clear skies indicate good visibility for navigation. The clean, well-maintained urban space suggests it's designed for pedestrian comfort and accessibility. This location seems ideal for

Description

"You are standing on a pedestrian street made out of red and black bricks.... On your right hand side, there is a large brick building, with glass windows at street level. The building has Gold letters that read Metro Credit Union. ... "

Sensitivity

Query 1 — Where is Metro Credit Union?

Answers
1-1. It is on your right hand side. **(True Positive)**
1-2. I do not know. **(False Negative)**

Specificity

Query 2 — Where is the nearest Starbucks?

Answers
2-1. I do not know. **(True Negative)**
2-2. It is in front of you. **(False Positive)**

Figure 2. Illustration of true-positives, true-negatives, false-positives, and false-negatives in responses. Description shown in the figure was written by humans.

strolling, with an inviting mix of historical and modern architecture.

The comparison between input formats of image only and AI-generated description can help determine optimum implementation method for future AI agents for BVI assistance. Technically, the AI agent could be implemented to generate directions directly based on interpretation of scene images, or first generate image caption and then generate directions based on the scene caption. The two implementation approaches may lead to different performances. The human-generated text input condition is included as a control condition to further evaluate the scene understanding capabilities of the AI agent, as compared with humans. This input format is not intended for practical use.

Responses from the various versions of the virtual assistants to each query was collected and evaluated for their accuracy in providing navigation guidance. The memory of the assistants' being evaluated was cleared between successive image inputs to avoid memorization of previous unrelated queries. The queries were submitted during July 2024.

Performance Evaluation

The ability of the virtual assistants to provide navigation assistance was evaluated based on their sensitivity (SEN) and specificity (SPE) in answering navigation queries. SEN was computed as the proportion of true positive to actual positive cases (answerable queries). SPE was the proportion of true negative to actual negative (unanswerable queries). Figure 2 shows examples of true-positive, true-negative, false-positive, and false-negative responses. In the case of answerable queries, true-positive responses were those that could be considered as consistent with the corresponding human-generated responses. Those inconsistent with human-generated responses (incorrect or made-up responses) were considered as false-negatives. In the case of unanswerable queries (when the requested information was not present in the scene image or associated textual description), the response "I do not know" or something to that effect was considered as true-negative, whereas any other responses (made-up) were considered as false -positives.

To judge the similarity of the responses from the navigation assistants and the human generated reference response (ground truth), we created a separate GPT-4o agent instructed to work as a grading assistant for comparing the two responses and providing a "yes" (similar answers or a match) or "no" (different answers) verdict. To validate the auto grading, both the AI grader and human graders adhered to the same evaluation criteria, which focused on whether the core information (i.e., navigation directions) in ChatGPT's responses aligned with the ground truth, instead of matching words or phrases directly. The 128 queries chosen for comparison of auto grading with manual grading came from 34 randomly chosen images from the dataset. A binary system was used for scoring the agreement between response and ground truth. Instructions emphasized matching based on overall consistency rather than matching exact words or phrases. Among the prompts we evaluated, the best prompt was "You will receive two sets of directions for a location, each in parenthesis and brackets. Your job is to determine whether they are roughly consistent by saying ‘Yes’ or ‘No.’ The directions don’t have to match word for word; the focus is on whether they lead to the same general direction or not.”

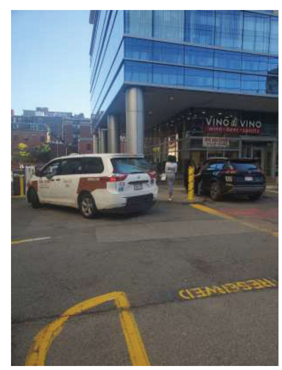
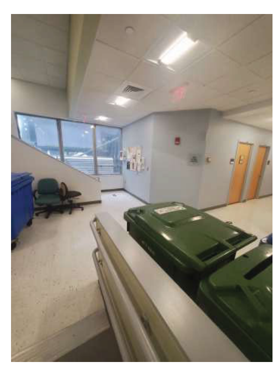
The grading assistant itself was evaluated by determining the degree of agreement between its judgment (whether ground truth and ChatGPT 4o responses are similar) and human judgment in a subset of queries (n = 128). Cohen’s Kappa coefficient was computed to determine the inter-rater reliability (between auto-grading and human grading), with the expectation that there would a high-level of consistency between the judgment of the virtual grading assistant and human grader, thereby indicating the validity of using the AI agent itself to determine the agreement between human generated and machine generated responses navigation queries. We further evaluated the effect of response length from ChatGPT 4o navigation assistants on the odds of agreement between the grading assistant and the human grader via a mixed-effects logistic regression model.

Results

Automated Grading of Responses

Of the 128 queries used for evaluating the agreement between the GPT-4o grading assistant and human grading, 22 were negative queries (17.2%) whereas the rest were positive queries. Although the GPT grading and human grading was not exactly the same (See examples in Table 3), there was a strong agreement between the grading assistant and the human grader on 114 responses (89%) – 91% agreement for negative queries and 88% on positive queries. Disagreements were almost evenly distributed across scenarios (3.9% for Streets, 2.3% for Indoor, Transit station and Plaza), and mostly occurred for Entrance destinations (5.5%), which was more than Direction (3.9%) and Specific objects (1.6%). The inter-rater reliability between grading assistant and the human grader was high (Cohen’s $\kappa = 0.72$, $z = 8.21$, $P < 0.001$). We invited another human grader (no. 2 or secondary grader) to judge the ground truth responses to determine the level

Table 3. Examples of Disagreements Between the GPT-4o Grading Assistant and the Human Grader. GPT Response Was Based on Image-Only Input Condition.

Query: Where is Vino di Vino? **Human Response:** Walk ahead and watch out for the moving cars. **GPT response:** Vino di Vino is located directly under the overhang of the modern building with large glass windows. You can identify it by its sign which is visible and has the text “Vino di Vino” along with “wine - beer – spirits” on it. The entrance to the store should be just beneath or adjacent to this sign. **GPT grader:** GPT response is correct. **Human grader:** GPT response is wrong, because it only described the location of the destination without providing directional guidance.	
Query: Where is Dr. Lee's office? **Human Response:** I don't know. There are two offices on your right. You can go to check if one of them is Dr. Lee's office. **GPT response:** There is no visible signage or information indicating the location of Dr. Lee's office. Typically, in a building like this, you would look for a directory or ask at a reception desk for specific office locations. If there's a floor plan or directory nearby, it would likely provide the necessary details to find Dr. Lee's office. Since I can't provide directions without additional information, I would suggest looking around for a building directory, following any additional signage, or asking a staff member for assistance. **GPT grader:** GPT response is wrong. **Human grader:** GPT response is correct, because it explicitly stated that there was no visual information about the destination.	

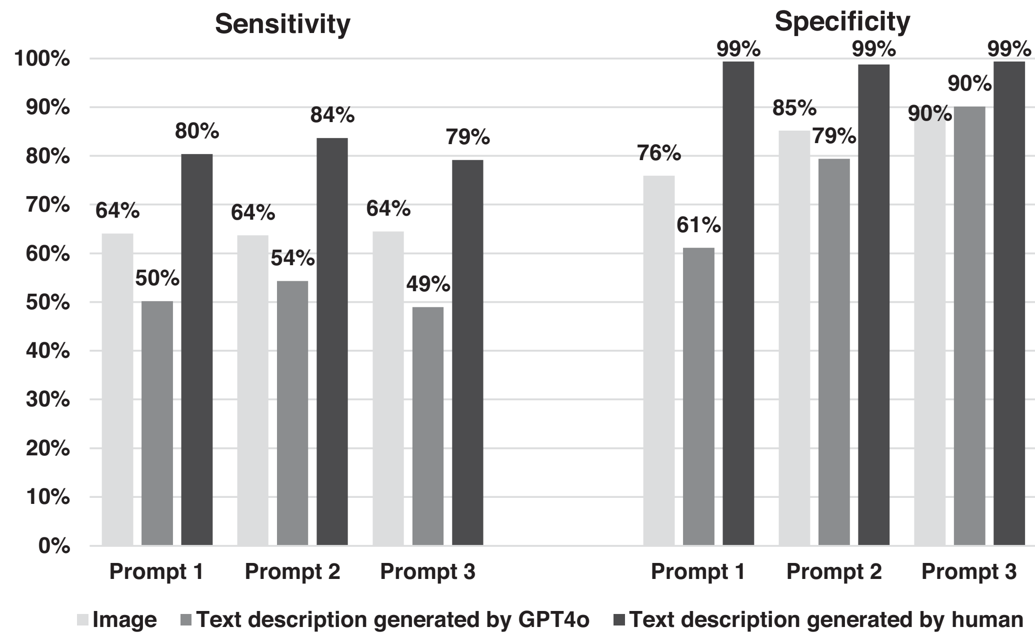


Figure 3. The SEN and SPE under different input configurations. The *white*, *gray*, and *dark bars* represent the input as image, text description generated by GPT-4o and text description generated by human, respectively. The left side group of bars is for SEN and the right side for SPE.

of agreement between the two human graders. The 2 human graders agreed on 107 responses (83%) with high inter-rater reliability (Cohen's $\kappa = 0.7$, $z = 8.21$, $P < 0.001$). Because the inter-rater reliability between automated grading assistant and the human grader was similar to that between the two human graders, auto grading using GPT-4o was considered as a valid grading assistant for judging whether the responses provided by navigation assistants were consistent with the human generated ground truth responses. There was no significant effect ($P = 0.48$) of ChatGPT 4o response length (number of words) on the odds of agreement with the ground truth response.

Accuracy of Responses to Navigation Queries

The responses of ChatGPT virtual assistants were verbose (Median [25th–75th percentile]: 34 [20–53] words) compared to the ground truth (9 [7–14] words; $P < 0.001$). For example, with Prompt 3, the micronavigation assistant gave the following responses to the two queries for the Street image shown in Figure 1.

Query 1: Where is Metro Credit Union? Response: "The Metro Credit Union is located on the right side of the image. It is in the brick building with the 'Metro Credit Union' sign clearly visible above the entrance."

Query 2: Where is the nearest Starbucks? Response: "The photo provided does not give any visible information about the location of a Starbucks. Therefore I cannot provide directions or confirm its proximity based on the current image."

The SPE and SEN with the three navigation prompts in three different conditions are shown in Figure 3. Worst SEN and SPE were obtained when GPT-4o was fed machine generated scene text description as input with Prompt 1 (without any additional instructions). With prompt 3 (the most detailed prompt), the SEN for GPT-4o generated text description did not improve, but SPE increased by around 30% points. SEN and SPE both improved about 13% and 7% points on average respectively, when GPT-4o answered queries directly based on the input images without the intermediate step of generating text description from input images. Furthermore, SEN and SPE both improved substantially, by about 17% and 16% points on average respectively, when human-generated text descriptions of the scenes were provided as input compared to image-only input. When the input was human generated scene text description, providing further prompt instructions to the assistants did not substantially change the SEN and SPE values. There was only a slight difference in SEN and SPE between indoor and outdoor scenarios (Table 4).

Table 4. Sensitivity and Specificity for Indoor and Outdoor Scenarios

Inputs	SEN	SPE
Image-only		
Outdoor	64%	87%
Indoor	64%	80%
Text description generated by ChatGPT 4o		
Outdoor	48%	78%
Indoor	54%	75%
Text description generated by human		
Outdoor	75%	100%
Indoor	88%	98%

Discussion

In this study, we investigated the ChatGPT 4o's ability to understand visual scenes and deliver natural language navigation instructions. Although the ultimate goal is to help BVI in real-world navigation, this study was focused on evaluating the current capabilities of AI agent to generate accurate responses to real-world navigation scenes. We compared the efficacy of the various combinations of inputs and tasks prompts to ChatGPT 4o in answering navigation-related queries, using a dataset of real-world navigation-related scenarios. In general, the navigation assistants provided the most accurate and specific responses when provided with human written description of the scenes instead of the scene image input directly, or the machine generated scene description. This indicates that the scene understanding capability of GPT-4o is somewhat limited, especially in the context of micronavigation. However, even with human generated text description of the scenes as input, the SEN on average was around 83%. The overall results indicate that the current general-purpose AI agents like ChatGPT 4o still have room for improvement and may not be optimally effective to work as micronavigation assistants to BVI individuals. Providing prompt-based instructions to the assistants mostly improved specificity of the responses. This indicates that even simple prompt-based instructions, can help in answering negative queries by reducing hallucination.

Despite not involving any BVI human subjects, this study is significant because it is the first time to investigate what could be expected from a general-purpose off-the-shelf AI agent in understanding navigation scenarios and providing useful response to navigation queries. LLM-based AI agents have already been released to public for assisting BVI (for example, BeMyAI function in BeMyEyes app). However, there has not been much published evidence regarding AI agents' utility for BVI users for navigation. The findings from this study can inform future study designs involving BVI users of AI technology for micronavigation. Navigation queries in our study did not require precise distance information, which we did not expect LLM to provide based on static images. The distance information, even if needed by the BVI for navigation, can be provided by complementary technologies, in combination with large language models to assist BVI.

For positive queries, the errors were mostly related to providing meaningful directions. First, the navigation assistants sometimes could not interpret images when the queried destination was related to directional signage. As the example in Figure 4A shows, ChatGPT instructed the user to go straight despite a left pointing arrow indicating the exit is to the left. Directional signage typically appears small in size in the scene images, which could further complicate the issue as the virtual assistants were unable to recognize small targets. Second, even though ChatGPT 4o could recognize the destination or target in the image, it sometimes gave vague answers without including necessary directions about how to get to the destination (Fig. 4B). Furthermore, GPT-4o might interpret the scene layout incorrectly in the text description, thus resulting in wrong directions (Fig. 4C). Less-than-perfect SPE is mainly caused by the hallucination of GPT. Providing prompts to the assistant helps to alleviate this issue to some extent.

Although the SEN is better with human written descriptions as input compared to image, it is still not close to 100%. Some of the errors seemed to be due to a combination of inference errors (incorrect scene understanding) and consequent hallucination of ChatGPT 4o. For example, when the scenario description "This is an arcade area. At the end of a straight passageway is a glass door entrance. There are glass windows on the left side of the passageway and at the end of the passageway. On the right side, there is a bike rack, a black trash bin, and a building column. A bike is positioned in the middle of the passageway. Beyond the right side of the passageway, there are three pedestrians: one at a medium distance and two further away. On the building, there is a sign that reads "Wholefoods-market"" was input into ChatGPT, and the query was: "Where is Wholefoods?" The response from ChatGPT was: "Wholefoods is located on the building that is to the right side of the passageway in the arcade area." This response suggests that ChatGPT did not understand the environment correctly and failed to infer that the building is not the right side. Hallucination phenomenon was also seen with text as input because ChatGPT considered the queried destination as related

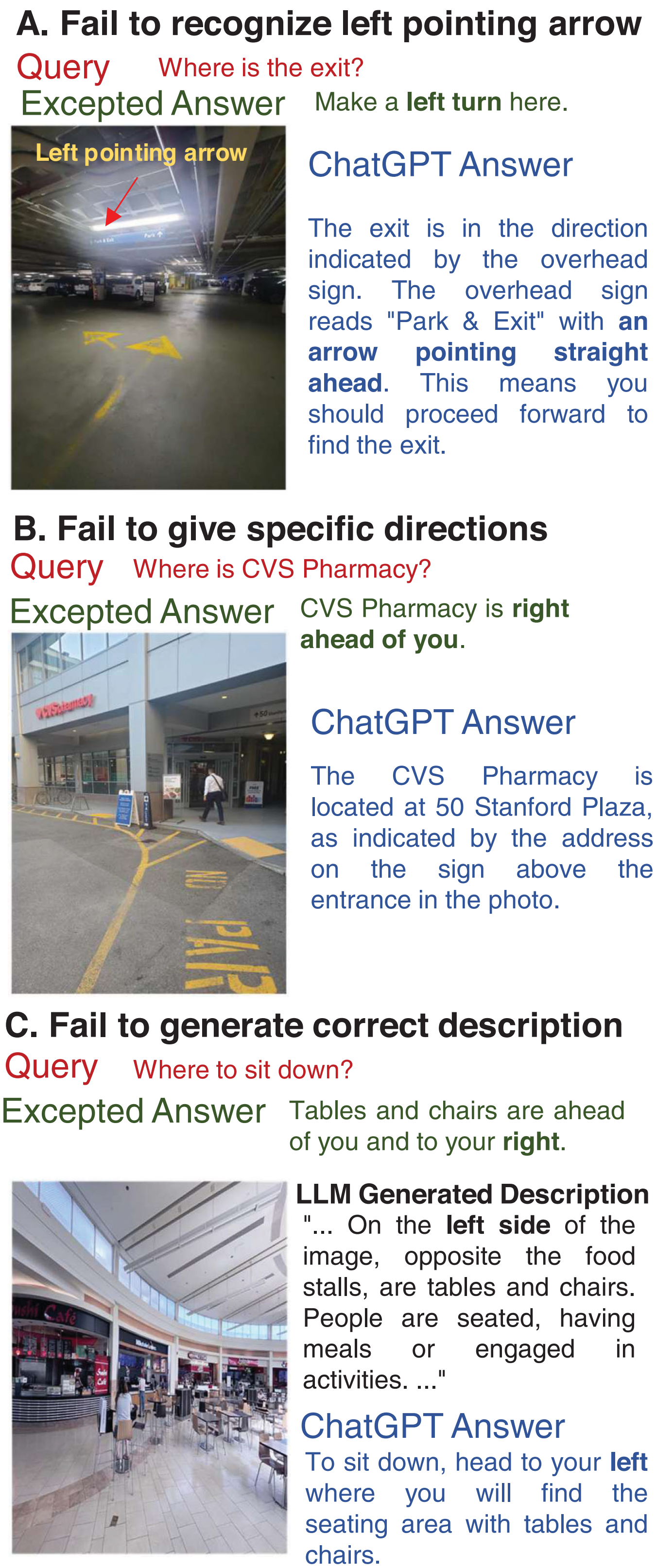


Figure 4. Examples of ChatGPT failed to give the correct directions.

to something mentioned in text descriptions, although there was no relationship between them. The scene description written by human did not include any information about where the building was, but the AI agent inferred that the building is to the right side of the passageway, probably because the scene description contained the statement: "beyond the right side of the passageway, there are three pedestrians." In the actual scene image, Wholefoods is not on the right side of passageway.

The reason why SPE was lower in AI-generated description condition than human-generated description condition is often due to hallucination, a common problem with LLM. AI-generated descriptions are much longer than human written descriptions. Long context might increase the chance of hallucination due to loss of focus in LLM.[28] Supplying good prompt helped improve the SPE of AI-generated description, as shown in our results (Fig. 3B).

The images included in this study were captured by sighted individuals and the quality could be different from pictures captured by people with vision impairments. It is likely that worse quality of input images would lead to further deterioration of AI model's performance. However, there are ways to mitigate the issue of poor image quality by employing one or a combination of different strategies including (1) a built-in quality check in the front end app that verifies basic image quality before inputting it to AI and suggests to the user to retake the picture, (2) training the users to appropriately use the app (via vision rehabilitation services/mobility specialists), (3) providing some indication to the users about device camera orientation based on the built-in motion sensor signal. For instance, the Camera app in iPhones provides aiming indications verbally when the voice over accessibility feature is turned on. However, such a human factor is out scope of this study.

Despite limitations of the current version of ChatGPT in providing navigation directions, BVI individuals could potentially benefit from advances in AI and multi-modal LLMs. A recent survey reported that ChatGPT has been rapidly adopted by BVI users in their daily life tasks, and the majority of those who were aware of it but not frequent users were interested in learning more about ChatGPT.[29] In the context of navigation, BVI travelers value the ability to improvise in unexpected situations. To date, this kind of task was only possible in the realm of human agents. But if AI virtual assistants could be proven to be adequate, then accessibility and quality of life of BVI could be improved. BVI users may require different levels of information for the same situation due to differences in vision status or based on personal preferences. In actual implementation, this can be done by setting specific prompts, for instance, "Limit response within 20 words," or letting the user ask follow-up questions to get more information from the scene as needed. Tailoring the response to suit user's preferences is future work.

Other datasets for navigation-related tasks exit, such as Talk2Nav[30] and R2R.[31] Talk2Nav is based on street view imagery (primarily for autonomous driving application) whereas R2R is for indoor navigation (e.g. living room to bedroom). Our dataset includes both indoor (such as shopping mall and underground subway station) and outdoor (most of them not on the street) images. When creating the dataset this study, an attempt was made to make it broadly representative of the daily-life navigation scenarios. It can be potentially used for comparing findings of future studies, either involving newer version of ChatGPT or other advances in foundational models in AI. R2R is somewhat related to our dataset and models trained using R2R dataset could potentially be further tuned for micronavigation assistance using our dataset. More scenarios can be added to the datasets to increase its diversity for future evaluations. Input from orientation and mobility specialists can be used to tune and improve the ground truth responses. The fast-evolving nature of AI models further highlights the time-sensitive nature of our research. The performance results may vary if a new testing dataset with significantly different characteristics is used, for instance, a dataset only focusing on public transportation.

According to our findings, the native ChatGPT 4o is still unable to provide correct micronavigation guidance in some cases, probably because its scene understanding is not optimized for navigation purposes. Given the currently limited capabilities of the AI agents, potential users could potentially use some strategies to boost performance for micronavigation. Based on our experimental results, using Prompt 3 from Table 2 leads slightly better outcomes. Further prompt engineering may be needed to extract more meaningful information while reducing verbosity. It is also likely that multiple pictures may need to be taken to complete the task step-wise, depending the complexity of the scene. How to interact with a virtual agent is still an active area of investigation in our opinion and more work needs to be done, especially via field-testing. The performance could be further improved by training custom VLMs for navigation-specific scenarios involving BVI travelers. Future work involves evaluating the utility of navigation directions provided by navigation assistants by BVI users.

Acknowledgments

We thank Prerana Shivshanker, Tony Liu, Jessie Zhou and Joye Luo for their assistance in creating the navigation scenario dataset and proofreading the paper.

Disclosure: **J. He**, None; **S. Pundlik**, None; **G. Luo**, None

References

1. Pundlik S, Shivshanker P, Luo G. Impact of apps as assistive devices for visually impaired persons. *Ann Rev Vis Sci.* 2023;9:12.11–12.20.
2. Parker AT, Swobodzinski M, Wright JD, Hansen K, Morton B, Schaller E. Wayfinding tools for people with visual impairments in real-world settings: a literature review of recent studies. *Front Educ.* 2021;6:723816.
3. Swobodzinski M, Parker AT. *A Comprehensive Examination of Electronic Wayfinding Technology for Visually Impaired Travelers in an Urban Environment: Final Report. NITC-RR-1177.* Portland, OR: Transportation Research and Education Center (TREC); 2019.
4. Crudden A. Transportation and vision loss: where are we now? *J Am Soc Ophthalmic Registered Nurs.* 2018;43(2):19–24.
5. Crudden A, McDonnall MC, Hierholzer A. Transportation: an electronic survey of persons who are blind or have low vision. *J Vis Impairment Blind.* 2015;109:445–456.
6. Luo G, Pundlik S. Widespread errors in bus stop location mapping is an accessibility barrier for passengers who are blind or have low vision. *J Vis Impairment Blind.* 2023.
7. Feng J, Beheshti M, Philipson M, Ramsaywack Y, Porfiri M, Rizzo JR. Commute booster: a mobile application for first/last mile and middle mile navigation support for people with blindness and low vision. *IEEE J Transl Eng Health Med.* 2023:523–535.
8. Pundlik S, Shivshanker P, Traut-Savino T, Luo G. Field evaluation of a mobile app for assisting blind and visually impaired travelers to find bus stops. *Transl Vis Sci Technol.* 2024;13(1):11–11.
9. Aira, https://aira.io/. Accessed November 22, 2024.
10. Nguyen BJ, Chen WS, Chen AJ, et al. Large-scale assessment of needs in low vision individuals using the Aira assistive technology. *Clin Ophthalmol.* 2019;13:1853–1868.

11. Nguyen BJ, Kim Y, Park K, et al. Improvement in patient-reported quality of life outcomes in severely visually impaired individuals using the AIRA assistive technology system. *Transl Vis Sci Technol.* 2018;7(5):30–30.
12. Park K, Kim Y, Nguyen BJ, Chen A, Chao DL. Quality of life assessment of severely visually impaired individuals using aira assistive technology system. *Trans Vis Sci Tech.* 2020;9(4):21.
13. Bordes F, Pang RY, Ajay A, et al. An introduction to vision-language modeling. *arXiv*, 2024, *arXiv:240517247*.
14. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. Presented at: *International Conference on Machine Learning*; 2021:8748–8763.
15. Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2017:3674–3683.
16. Hermann KM, Malinowski M, Mirowski PW, Banki-Horvath A, Anderson K, Hadsell R. Learning to follow directions in street view. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020;34:11773–11781.
17. Huang C, Mees O, Zeng A, Burgard W. Visual Language Maps for Robot Navigation. *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2022:10608–10615.
18. Liang X, Zhu F, Lingling L, Xu H, Liang X. *Visual-language navigation pretraining via prompt-based environmental self-exploration*. arXiv, 2022, arXiv:2203.04006.
19. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. *Adv Neural Inform Process Syst*. 2023;36:34892–34916.
20. Moudgil A, Majumdar A, Agrawal H, Lee S, Batra D. Soat: a scene- and object-aware transformer for vision-and-language navigation. *Adv Neural Inform Process Syst*. 2021;34:7357–7367.
21. Shah D, Osinski B, Ichter B, Levine S. LM-Nav: robotic navigation with large pre-trained models of language, vision, and action. *6th Annual Conference on Robot Learning*; 2023:492–504.
22. Tsai Y-HH, Dhar V, Li J, Zhang B, Zhang J. Multimodal large language model for visual navigation. arXiv, 2023, arXiv:231008669.
23. Vasudevan AB, Dai D, Van Gool L. Talk2Nav: Long-Range Vision-and-Language Navigation with Dual Attention and Spatial Memory. *Int J Comput Vis.* 2021;129:246–266.
24. Yu B, Kasaei H, Cao M. L3MVN: Leveraging Large Language Models for Visual Target Navigation. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2023.
25. Zhu F, Liang X, Zhu Y, Chang X, Liang X. SOON: Scenario Oriented Object Navigation with Graph-based Exploration. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021:12684–12694.
26. OpenAI. GPT4 Vision API Guide. Available at: https://platform.openai.com/docs/guides/vision. Accessed November 22, 2024.
27. OpenAI. Available at: https://help.openai.com/en/collections/3675942-prompt-engineering. Accessed November 22, 2024.
28. Wu Z, Liu B, Yan R, Chen L, Delteil T. Reducing Distraction in Long-Context Language Models by Focused Learning. arXiv, 2024, arXiv:24110 5928.
29. Pundlik S, Shrivshanker P, Bowers AR, Luo G. Use of ChatGPT among blind and visually impaired individuals. *J Vis Impairment Blind.* 2024;118:118–119.
30. Vasudevan A, Dai D, Gool L. Talk2Nav: long-range vision-and-language navigation with dual attention and spatial memory. *Int J Comput Vis.* 2021;129:246–266.
31. Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018:3674–3683.

Appendix I

Code (1)

```
from openai import OpenAI

client = OpenAI()

assistant = client.beta.assistants.create(

 name=" Simple-Navigation",

 instructions=" You are a navigator that should give correct directions to
help a person go to where they need to go.",

model="gpt-4o",

)
```

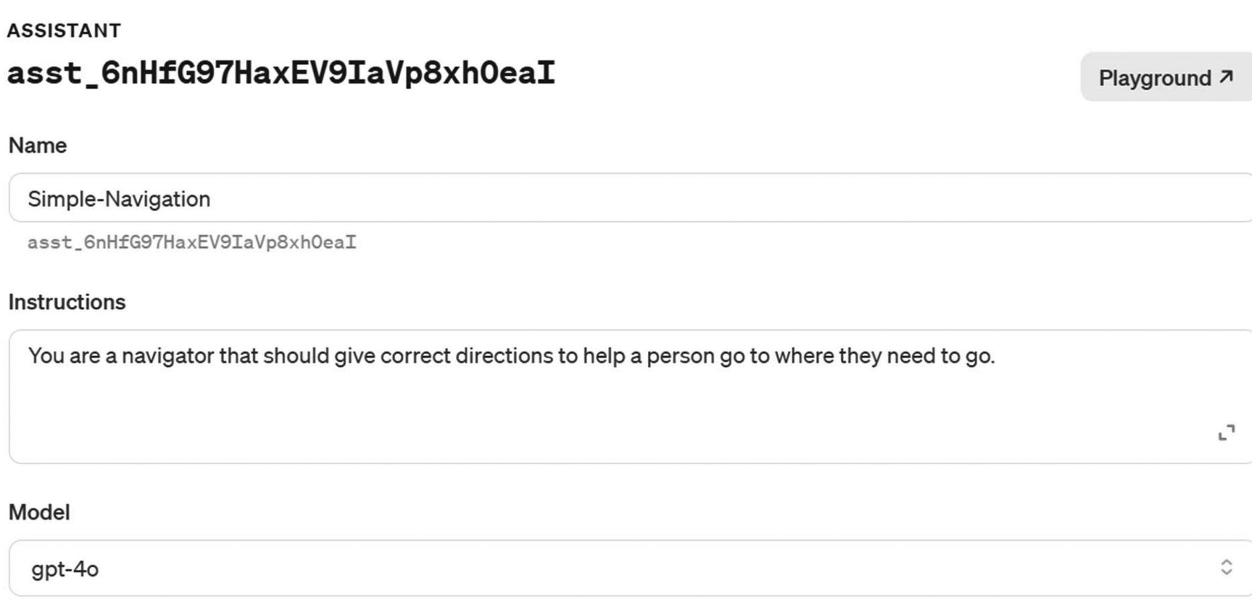


Figure A1. Ways of providing prompts to ChatGPT bots: via “instructions” parameter in code (top) or by “Instructions” text box in UI.

translational vision science & technology