Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/81/20241085

SOVAR: System of Visual Assistance and Recognition

Ken Huang

The Athenian School, Danville, USA

kenhuangsy@gmail.com

Abstract. Around 295 million individuals globally suffer from moderate to severe vision impairment. They struggle with daily activities and depend heavily on others for assistance. Leveraging augmented reality (AR) and artificial intelligence (AI), we have developed SOVAR, a mobile application that enables greater independence for visually impaired individuals in their daily lives. SOVAR includes two modules: navigation and scene understanding. Navigation involves two phases: mapping and guidance. During mapping, SOVAR builds and optimizes the maps with key locations labeled by users via voice input. During guidance, SOVAR plans the path and guides users to requested key locations, with visual and audio assistance and realtime obstacle avoidance. The scene understanding module includes a Large Vision Language Model (LVLM) to help the users through image captioning and visual question answering. For navigation, our user study shows that participants successfully navigated to the key locations from three separate locations in 86.67% of trials without intervention. The success rate improves with increased user familiarity with the application. For scene understanding, our study shows that leveraging LVLMs to help the visually impaired allowed the participants to answer the visual-related questions with an accuracy of 100%. In this work, we developed SOVAR, a mobile application leveraging AR and AI to assist the visually impaired in navigation and scene understanding. The promising results from the user study on SOVAR demonstrate the effectiveness of AR and AI for visual assistance and indicate their potential impact on general assistive technologies.

Keywords: Augmented reality, AI, navigation, visually impaired.

1. Introduction

The eye is one of the most critical parts of the human body, acting as the primary gateway to the world around us. It allows us to interpret our surroundings, understand nuances, and navigate through our day-to-day lives. However, for the visually impaired community, this gateway is obscured or is lost entirely, preventing them from a seamless interaction with the world. According to the World Health Organization (WHO), at least 2.2 billion people globally are facing visual impairment [1]. According to George- town University, in the United States, almost 20 million Americans, which is about 8% of the population, have visual impairments [2]. According to the same source, visual impairment limits social interactions and leads to higher rates of depression, higher healthcare usage, and lower earnings.

Individuals with visual impairments face many daily challenges, ranging from simple tasks like finding objects at home to more complex problems in public spaces. For them, locating any object can become a daunting task, and public areas present even greater difficulties. Grocery shopping can be cumbersome without the ability to see price tags, and handling money is problematic as they cannot

DOI: 10.54254/2755-2721/81/20241085

differentiate between currency denominations. Transportation and navigation also present significant hurdles, such as identifying traffic signal colors and differentiating between obstacles. These common issues affect both personal tasks and public interactions, hindering independence and complicating everyday life.

Each of these examples underlines the everyday challenges faced by the visually impaired and underscores the necessity for innovative solutions. It is within this context that our research, aimed at leveraging the fast development of multi-modal systems, large language models (LLMs), and augmented reality (AR) to assist individuals with visual impairments, becomes significant.

In recent months, the field of artificial intelligence (AI) has witnessed a remarkable surge in the development and advancement of LLMs. Pioneering LLMs such as ChatGPT[3] have captured global attention, demonstrating unprecedented capabilities in understanding and interpreting human language. These models are distinguished by their ability to process complex human inputs and generate nuanced, in-depth responses. In the context of assisting the visually impaired, LLMs present a new approach to interpreting complex user queries. They can discern the underlying needs and intentions behind language, translating them into actions. We leverage LLMs in our system to discern the user's needs based on their language input and select the most suitable tool to help them. Along with LLMs, there has been significant progress in the development of multi-modal models, specifically Large Vision Language Models (LVLMs), which utilize LLMs to generate remarkable responses to queries with images. LVLMs offer a new approach to addressing the unique challenges faced by the visually impaired, leveraging the strengths of both visual interpretation and textual communication to provide scene understanding assistance.

AR has also made significant progress in recent years, evolving from a niche technology to one with widespread applications across many industries. AR-capable devices like smartphones, tablets, and smart glasses have become more powerful, with better sensors, displays, and processing capabilities. This has enabled more sophisticated and immersive AR experiences. A prime example is the Apple Vision Pro released in early 2024, which features dual micro-OLED displays with a combined 23 million pixels, providing incredibly sharp and immersive visuals, eye tracking, hand gestures, and voice commands [4]. The AR market is projected to grow significantly with estimates suggesting it could reach $1,869.40 billion by 2032 [5].

We introduce SOVAR, a mobile application leveraging AR and AI to assist the visually impaired in navigation and scene understanding. We introduce a voice command and assistive module selection system, leveraging LLMs, that allows our system to cater to the needs of individuals with varying degrees of visual impairment. Since our system is based on a model selection process, it is adaptable and flexible. New assistive modules can be added to the system to better serve the visually impaired. The promising results from the user study on SOVAR demonstrate the effectiveness of AR and AI for visual assistance and indicate their potential impact on general assistive technologies.

2. Related Works

Vision-assistive frameworks: Many vision-assistive frameworks have been developed to aid the visually impaired in navigating outdoor and indoor environments. For example, Kim et. al. introduced a robot guide dog to help the visually impaired navigate [6]. Currently, there are apps developed to aid the visually impaired to better understand their surroundings and seek help from volunteers, such as Microsoft's Seeing AI and Be My Eyes. However, the interface and methods used in these apps are not user-friendly for those who are blind and do not leverage the fast-advancing developments in the AI space, causing their performance and their ability to help the visually impaired to be limited. This underlines the significance of a new system that can leverage audio interaction and the newest advancements.

Multi-modal Models: In recent months, there has been a considerable rise in the adoption of Large Language Models (LLMs) for various tasks, both in the realm of language and beyond [7][8]. Notably, the application of LLMs has been particularly effective in vision-related tasks, signaling a significant advancement in this field [9][10][11][12]. These systems, known as Large Vision Language Models

DOI: 10.54254/2755-2721/81/20241085

(LVLMs), are capable of generalizing across a variety of domains. This range includes but is not limited to, Visual Question Answering (VQA), image captioning, and semantic segmentation. However, in the scope of this paper, our focus will be directed specifically toward VQA. VQA is a task in computer vision where the model is presented with a user's query in natural language concerning the content of a specific image, and it must then generate a pertinent response that aligns with the visual information. The improvements in VQA brought on by LVLMs could drastically enhance the quality of life and autonomy for these individuals.

Multi-Modal Systems: In addition to multi-modal models, the rise in popularity of multi-modal systems is also worthy of note [13][14]15]. These innovative systems are designed with the overarching goal of leveraging the inherent capabilities of LLMs to drive decision-making processes. More specifically, they utilize LLMs as the central processing unit that, based on user input, discerns which tool or resource would be the most suitable for fulfilling the user's request. The toolkit these systems can access is vast, encompassing resources such as object detectors, video narrators, subtitle grounding mechanisms, and more. These tools serve a variety of functions, each playing a unique role in aiding the completion of a user's query. In alignment with this methodology, our research also adopts a similar approach. We use an LLM as the core processing unit that, upon receiving a user's query, determines the most appropriate assistive module to use. We built our assistive modules, that are specifically tailored to assisting the visually impaired.

3. Methods

3.1. Leveraging LLMs for Assistive Module Selection

The central thesis in the paper HuggingGPT [15] postulates: "Language is a generic interface for LLMs to connect with AI models." This principle is embodied in our design architecture (See Figure 1) where we leverage LLMs to serve as an intermediary between user queries and the selection of tools. After a user records a query in the mobile app, we first transcribe this query into text with the OpenAI Whisper model [16]. This conversion is necessary because LLMs operate on text, not audio. To reduce response latency, we utilize Groq's Whisper API, which offers significantly faster performance. It operates at 164 times real-time speed, transcribing 164 seconds of audio in just 1 second of processing time [17]. Once we have a textual representation of the query, we feed it to a tool-calling Langchain LLM agent [18]. A LLM agent essentially uses the LLM as a reasoning engine to determine which actions to take and in which order. We refer to the LLM agent as the "controller" in our system. The controller's role is to analyze the user query and subsequently identify the appropriate assistive tool, called "executor," to use based on our description of each tool. This tool selection architecture is adaptable, as new "expert" assistive tools can be integrated into the system for specific tasks.

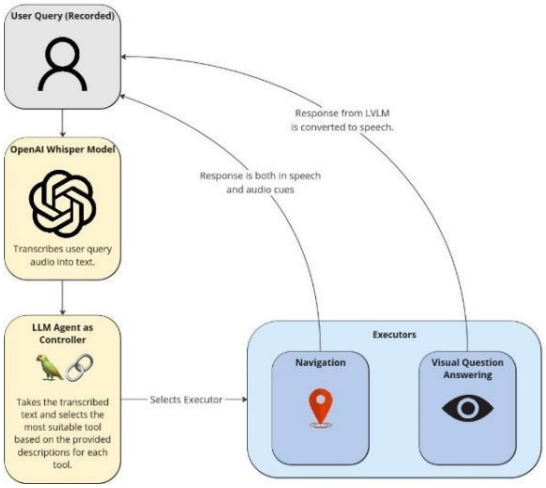

Figure 1. Language Powered System Overview

DOI: 10.54254/2755-2721/81/20241085

3.2. Navigation

To help the users with navigation, we use Apple's AR and computer vision frameworks to help with mapping the environment, localizing the user, guiding the user, and avoiding obstacles. Navigation is divided into two phases: mapping and guiding.

3.2.1. Mapping the Environment. In the mapping phase, users interact with the app to create a network of waypoints. By clicking on any horizontal surface (such as a floor), users can place waypoints. Subsequent clicks on other points within the same plane automatically connect these waypoints. Users can assign optional names to waypoints (e.g., kitchen, bathroom) through a long-press gesture (see Fig. 2. Clicking on an existing waypoint highlights it and clicking on another existing waypoint will connect the two waypoints if they were not previously connected. If two waypoints were already connected, clicking on both of them will remove the connection between them. Once the waypoint network is complete, users can name the entire map (e.g., gym, house, office) and submit it to the backend. The backend processes this information as a bidirectional graph, which is then stored in a Supabase database (see Fig. 3). This database utilizes pgvector to embed the map's name and waypoint information. The ARWorldMap [19], which is an object that contains a snapshot of all the spatial information used by ARKit to locate the user's device in real-world space is serialized into a byte array and saved into the database. This allows the AR world mapping data to be restored when the user loads a map and localize the user within the context of the map.

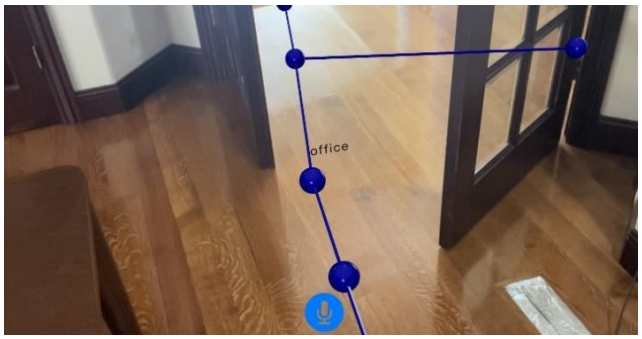

Figure 2. Connected waypoints and a waypoint with a name

For efficient retrieval and processing, we implement a hybrid search approach [20]. This method combines full-text search (searching by keyword) with semantic search (searching by meaning) to identify results that are both directly and contextually relevant to the user's query. The full-text search capability allows users to find maps and waypoints during the guiding phase based on exact keyword matches, while the semantic search leverages the embedded vector representations to find conceptually similar results, even when the exact keywords aren't present. For instance, if the user wants to use the "office" map but they accidentally said "workplace," the app can still find the "office" map.

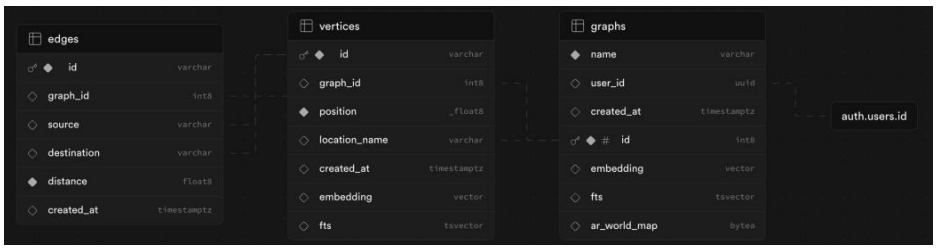

Figure 3. Database Schema

3.2.2. Guiding. In the guiding phase, users can initiate navigation by recording an audio query specifying their desired destination on a particular map (e.g., "Navigate me to my desk in the office"). The LLM agent processes this query, identifying the navigation intent and extracting both the destination

DOI: 10.54254/2755-2721/81/20241085

and map names. The system then employs a two-step hybrid search process: first, to identify the most relevant map, and second, to locate the closest matching waypoint within that map. Once identified, the backend loads the entire map structure into a bidirectional graph. The user's current position is dynamically connected to the nearest node in this graph based on position. Finally, the A* pathfinding algorithm computes the optimal route from the user's position to the target waypoint. Audio cues are used to help navigate the user to the destination by continuously guiding the user to the next closest node in the graph.

3.2.3. Audio Cues for Navigation. SOVAR uses audio cues generated using a combination of frequency modulation synthesis and spatial audio techniques to provide intuitive directional and distance information.

3.2.3.1. Frequency Modulation Synthesis. At the core of the audio cue system is a real-time FM synthesis engine. FM synthesis produces complex audio waveforms by modulating a carrier signal's frequency with a modulator signal. The synthesizer is defined by the following equation:

$$s(t) = sin\ (2\pi f_c t + A_m sin\ (2\pi f_m t)) \tag{1}$$

where:

- $s(t)$ is the output signal
- f_c is the carrier frequency
- f_m is the modulator frequency
- A_m is the modulator amplitude
- t is time

The synthesizer operates at a sample rate of 44.1 kHz, producing high-quality audio output. It utilizes a circular buffer system with two buffers, each containing 1024 samples, to ensure low-latency playback while maintaining efficient CPU usage.

3.2.3.2. Distance Encoding

The distance to the next waypoint is encoded in the frequency of the audio cue. This mapping is implemented as follows:

$$f = f_{min} + (f_{max} - f_{min})(1 - \frac{min(d, d_{max})}{d_{max}}) \tag{2}$$

where:

- f is the output frequency
- f_{min} is the minimum frequency (220 Hz, A3 note)
- f_{max} is the maximum frequency (880 Hz, A5 note)
- d is the distance to the waypoint
- d_{max} is the maximum considered distance (10 meters)

This logarithmic mapping ensures that smaller distances result in higher pitches, providing an intuitive "closer means higher" association for users.

3.2.3.3. Directional Cues

Directional information is conveyed through stereo balance and verbal cues. The balance of the audio signal is calculated based on the angle between the user's current heading and the direction to the next waypoint:

$$b = sin\ (\theta \frac{\pi}{180}) \tag{3}$$

where:

DOI: 10.54254/2755-2721/81/20241085

- b is the stereo balance (-1 for full left, 1 for full right)
- θ is the angle between the user's heading and the waypoint direction

This creates a natural "sound stage" effect, where the audio cue appears to come from the direction of the waypoint relative to the user's current orientation. Additionally, verbal cues are provided when significant changes in direction are required. The system monitors the angle to the waypoint and triggers spoken directions ("Turn left", "Turn right", or "Go straight") when the angle changes by more than 45 degrees. This hybrid approach of continuous tones and discrete verbal cues has been shown to be effective in navigation tasks for visually impaired individuals.

3.2.3.4. Temporal Characteristics

The frequency of audio cue playback is dynamically adjusted based on the distance to the waypoint. This is achieved by varying the interval between cues:

$$I = I_{min} + (I_{max} - I_{min}) \frac{min\,(d, d_{max})}{d_{max}} \tag{4}$$

where:

- I is the interval between cues
- I_{min} is the minimum interval (0.1 seconds)
- I_{max} is the maximum interval (2.0 seconds)
- d is the distance to the waypoint
- d_{max} is the maximum considered distance (10 meters)

This approach ensures that cues are played more frequently as the user approaches a waypoint, providing a sense of urgency and increased precision for final approach navigation.

3.2.3.5. Obstacle Detection

In addition to waypoint navigation, the system provides audio cues for obstacle detection, specifically for doors. When a door is detected within 0.5 meters of the user's current position and along their path to the next waypoint, the system announces "Door ahead" using text-to-speech synthesis. This feature helps users anticipate and navigate doorways safely.

3.2.3.6. Implementation Details

The audio cue system is implemented using Apple's AVAudioEngine framework, which provides low-latency audio playback and processing capabilities. The system uses a combination of Swift's Grand Central Dispatch (GCD) for concurrent audio generation and Core Motion framework for device orientation tracking. This ensures smooth, real-time audio feedback even on mobile devices with limited processing power.

3.2.4. Visual Question Answering (VQA). To assist users with visual-related queries about the scene in front of them, SOVAR utilizes OpenAI's GPT-4o API, which includes advanced vision capabilities (see 4). The user's query, along with a snapshot of the scene, is provided to GPT-4o. The output is then conveyed to the user through text-to-speech functionality.

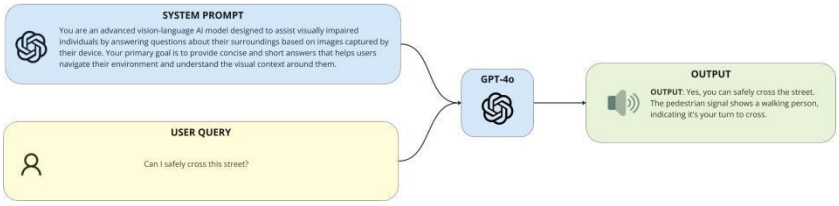

Figure 4. VQA

DOI: 10.54254/2755-2721/81/20241085

4. User Study

The navigation module was evaluated through a comprehensive user study involving five blindfolded subjects. Each participant was equipped with a white cane and tasked with navigating to three distinct locations across three separate trials. The trials were designed to test various navigation scenarios:

- Multi-level navigation (including ascending and descending stairs)
- Single-floor navigation between rooms
- Transition from indoor to outdoor environments

Subjects used voice input to request navigation instructions to each destination. A trial was deemed unsuccessful if researcher intervention was necessary to prevent potential harm to the participant.

The results of the navigation tests are presented in Table 1. The data indicates that SOVAR demonstrates high capability in assisting with navigation tasks. Initially, some users exhibited unfamiliarity with the application's interface, causing the low success rate for the first trial. However, after receiving appropriate guidance and instruction, participants quickly adapted to the system's functionality. It is worth noting that the learning curve for SOVAR appears to be relatively short, suggesting that with minimal training, users can effectively leverage the system for navigation purposes. This rapid adaptability is a promising indicator of SOVAR's potential for real-world applications in assistive technology for visually impaired individuals.

Table 1. Navigation Test Results

Navigation	Trial 1	Trial 2	Trial 3
Success rate without intervention (%)	60	100	100

5. Conclusion

This research presented SOVAR, an innovative multi-modal assistive system designed to aid the visually impaired with navigation and scene understanding through natural language interactions. We introduced a flexible framework centered around an LLM controller that analyzes user queries and selects appropriate modules to address their visual assistance needsstanding through natural language interactions. We introduced a flexible framework centered around an LLM controller that analyzes user queries and selects appropriate modules to address their visual assistance needs.

The navigation module of SOVAR demonstrated promising results in our user study, with an overall success rate of 86.67% across various navigation scenarios. The rapid learning curve observed suggests that users can quickly adapt to and effectively utilize the system with minimal training. This adaptability, combined with the system's ability to handle complex navigation tasks such as multi-level navigation and indoor-outdoor transitions, highlights SOVAR's potential for real-world applications.

The integration of advanced technologies such as AR, AI, and LVLMs in SOVAR represents a significant step forward in assistive technology for the visually impaired. By leveraging these technologies, SOVAR offers a more intuitive and comprehensive solution compared to existing assistive apps, addressing both navigation and scene understanding challenges.

The modular design of SOVAR, with its LLM-powered tool selection mechanism, ensures that the system is adaptable and extensible. This design choice allows for the easy integration of new assistive modules as technology advances, ensuring that SOVAR can evolve to meet the changing needs of its users.

While our initial results are promising, further research is needed to evaluate SOVAR's long-term effectiveness and usability in diverse real-world scenarios. Future work should focus on:

1. Expanding the user study to include a larger and more diverse group of visually impaired individuals.
2. Enhancing the system's obstacle avoidance system to help the user navigate without white canes.
3. Improving the VQA module to provide more detailed and contextaware scene descriptions.
4. Exploring the integration of haptic feedback to complement audio cues for navigation.

DOI: 10.54254/2755-2721/81/20241085

5. Investigating the potential of SOVAR in other assistive technology applications beyond visual impairment.

In conclusion, SOVAR demonstrates the significant potential of integrating AR, AI, and natural language processing in creating more effective and user-friendly assistive technologies. By addressing both navigation and scene understanding challenges, SOVAR takes a crucial step towards enhancing the independence and quality of life for individuals with visual impairments. As we continue to refine and expand this system, we envision a future where technology can more seamlessly bridge the gap between visual impairment and environmental interaction, fostering greater autonomy and inclusivity for all.

References

[1] World Health Organization. (n.d.). *Blindness and visual impairment*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (Accessed on August 14, 2024).

[2] Georgetown University. (n.d.). *Visualizing health policy*. Georgetown University Health Policy Institute. https://hpi.georgetown.edu/visual/ (Accessed on August 14, 2024).

[3] OpenAI. (n.d.). *ChatGPT*. OpenAI. https://chat.openai.com/ (Accessed on August 14, 2024).

[4] Apple Inc. (n.d.). *Apple Vision Pro*. Apple. https://www.apple.com/apple-vision-pro (Accessed on August 14, 2024).

[5] Fortune Business Insights. (n.d.). *Augmented reality (AR) market size, share & COVID-19 impact analysis, by component, by device type, by industry, and regional forecast, 2023-2030.* Fortune Business Insights. https://www.fortunebusinessinsights.com/augmented-reality-ar-market-102553 (Accessed on August 14, 2024).

[6] Kim J T, Yu W, Kothari Y, et al. Transforming a quadruped into a guide robot for the visually impaired: Formalizing wayfinding, interaction modeling, and safety mechanism[J]. arXiv preprint arXiv:2306.14055, 2023.

[7] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 3(6):7, 2023.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth´ee Lacroix, Baptiste Rozi`ere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[12] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. arXiv preprint arXiv:2306.17107, 2023.

[13] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. arXiv preprint arXiv:2306.08640, 2023.

[14] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Internchat: Solving vision-centric tasks by interacting with chat- bots beyond language. arXiv preprint arXiv:2305.05662, 2023.

DOI: 10.54254/2755-2721/81/20241085

[15] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580, 2023.
[16] Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//International conference on machine learning. PMLR, 2023: 28492-28518.
[17] Groq. (n.d.). *Groq runs Whisper Large v3 at a 164x speed factor according to new Artificial Analysis benchmark*. Groq. https://wow.groq.com/groq-runs-whisper-large-v3-at-a-164x-speed-factor-according-to-new-artificial-analysis-benchmark/ (Accessed on August 14, 2024).
[18] LangChain. (n.d.). *Agent types*. LangChain Documentation. https://python.langchain.com/v0.1/docs/modules/agents/agent_types/ (Accessed on August 14, 2024).
[19] Apple Inc. (n.d.). *ARWorldMap*. Apple Developer Documentation. https://developer.apple.com/documentation/arkit/arworldmap (Accessed on August 14, 2024).
[20] Supabase. (n.d.). *Hybrid search*. Supabase Documentation. https://supabase.com/docs/guides/ai/hybrid-search (Accessed on September 19, 2024).