Coding Quiz: App Development with Multimodal Interaction

Introduction

The UHN AI Hub is a collaborative center dedicated to enhancing human intelligence through healthcare innovation. By advancing AI technologies and accelerating their application, the Hub aims to deliver the highest standard of patient care while empowering healthcare providers. This quiz is designed for applicants seeking a software developer position within the UHN AI Hub. While candidates are not expected to develop novel AI models, they should demonstrate the ability to efficiently and securely deploy state-of-the-art AI models, ensuring they are accessible to end-users such as clinicians and patients.

Objective

Develop a mobile or web application that integrates multimodal interaction functionalities for image analysis, focusing on usability, robustness, and seamless functionality. Here is a demo from Molmo.

- input: image + audio or text
- output: audio
- key function: call out-of-the-box API or open-source models to
    1. convert user voice input requirements (e.g., please describe the image?) to text (e.g., OpenAI whisper)
    2. parse the text and image by vision-language models (e.g., OpenAI GPT-4 or Qwen2-VL)
    3. convert the text output to audio by text-to-speech models

The following requirements must be addressed:

Core Development Requirements

1. **Implement the following features:**
    - Audio recording functionality to capture user voice input.
    - Camera access for users to take photos.
    - Multimodal data processing to parse text, audio, and image inputs
    - Text-to-speech capability to convert and play the processed text as audio.
2. **UI Components:**
    - A text input box
    - A button to trigger audio recording.
    - A button for photo capture.
    - A window to show processed results.

Advanced Features (Optional for Extra Scores)

- Address potential exceptions during multimodal data processing, such as invalid input, API timeouts, or camera access issues.
- Efficient resource usage on varying device capabilities.
- Responsive design to support different screen sizes and orientations.

Submission

- GitHub link to your code. Please provide a well-structured ReadMe file.
- Google Drive download link to the demo video: Record a video highlighting the core features and seamless interactions within the application.

Assessment Criteria

- **Feature Implementation:** Accuracy and completeness of the implemented features.
- **Code Quality:** Readability and maintainability of the code.
- **Design:** Usability and user experience of the interface.
- **Performance:** Responsiveness and optimization across devices.
- **Robustness:** Thoroughness of error handling and resilience of the application.
- **Documentation:** Quality and comprehensiveness of the README file and demonstration video.