

# See the World Through LLMs

LAI

[pubudu1ekanyake@gmail.com](mailto:pubudu1ekanyake@gmail.com)

# Problem Definition

It is quite difficult to make the blind people see the world even though there are developments in the technologies or it is difficult to economically feasible to provide these technologies to majority of the blind people.

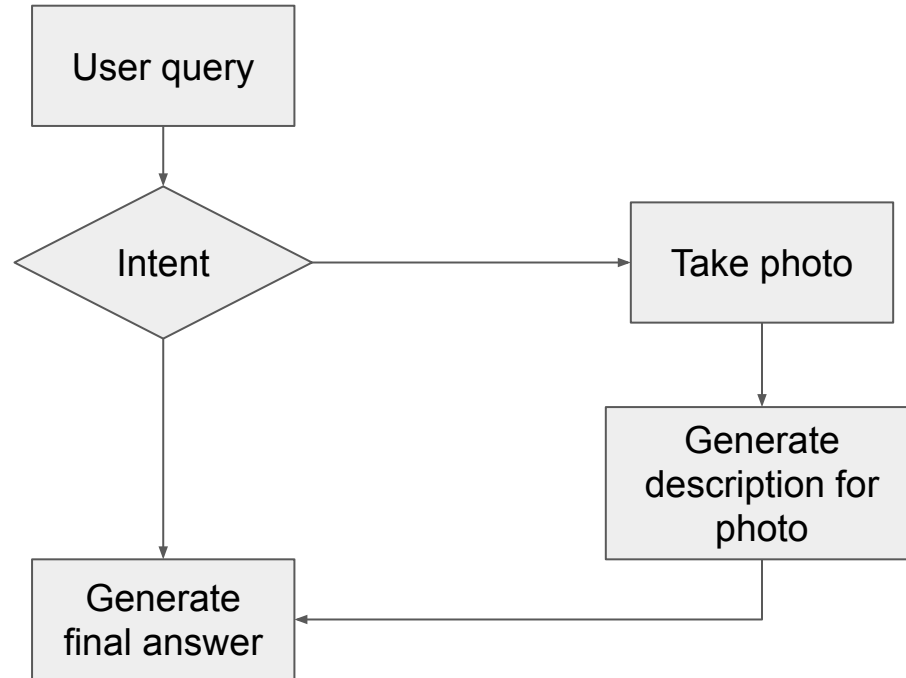
# Solution and Proposal

With development of the LLMs and related technologies, I see that we can reduce the gap of inability to afford the available products for the blind people to see the world.

With api based llms, we can provide chatbots to assist blind people to not only see the world but also to make their day today lives easier.

Start of my project I will use google gemini-pro and google gemini-pro-vision as the llms and midas model to estimate and get the depth map image.

# Methodology - Flow



- User enters the query.
- If the user wants to know about his/her surrounding then need to take a photo
- If not for the asked question the llm will directly generate the answer
- If photo is required to be taken the photo will be used to generate the depth map image.
- Using the query, photo and depth map image, we generate a description about the surrounding/view
- Finally generate the final answer based on the query and generated description

# Sample Chat Conversation



Next slide we will provide a conversation carried out based on given image here.

Using cache found in /home/aida/learn/colin/hub/inter-1st-1800-master

User ('q' to exit): hello

No photo required

Chatbot: Hello, how can I assist you today?

User ('q' to exit): is there any shop near me that i can buy food

Photo is required

Chatbot: I am not sure. The context does not mention anything about shops or food.

User ('q' to exit): ok explain the surrounding of me please?

Photo is required

Chatbot: The context describes a street scene in London, England. The street is lined with tall buildings, many of which are covered in scaffolding. There are a number of red double-decker buses on the street, as well as a few cars. The street is wet from the rain, and the sky is overcast. The approximate distance to the nearest bus is 4 meters. The approximate distance to the buildings is 15 meters. The climate is temperate, with mild summers and cool winters.

User ('q' to exit): ok, now tell me is there any shops that i can buy food?

Photo is required

Chatbot: Yes, there are several shops that you can buy food from. For example, you can go to the Pret a Manger at 96 Regent Street or the EAT at 86-90 Regent Street.

User ('q' to exit): is it in a visible range of my view?

No photo required

Chatbot: The context does not mention anything about your view, so I cannot answer this question from the provided context.

User ('q' to exit): ok is the mentioned shop near me?

Photo is required

Chatbot: No, the mentioned shop is not near you.

User ('q' to exit): what is the mentioned shop?

No photo required

Chatbot: The mentioned shop is Pret a Manger at 96 Regent Street or the EAT at 86-90 Regent Street.

User ('q' to exit):

Please note that 'No photo required' and 'Photo is required' are the two intents.

# Possible Improvements

- This project is a starter point for the project
- Possible improvements:
  - Optimize the code improve with an UI interface (if required) or device (which is preferable)
  - Add voice to text and text to voice features.
  - Currently we use still images but required to add instant photo taking capability (+ video)
  - Improved prompts
  - Improved depth estimation model (stereo images)
  - Adding guardrails for proper guiding
- Please note that the image used in the slide #5 is taken from the internet and full credit belongs to the one who took that photo.

Thank You.