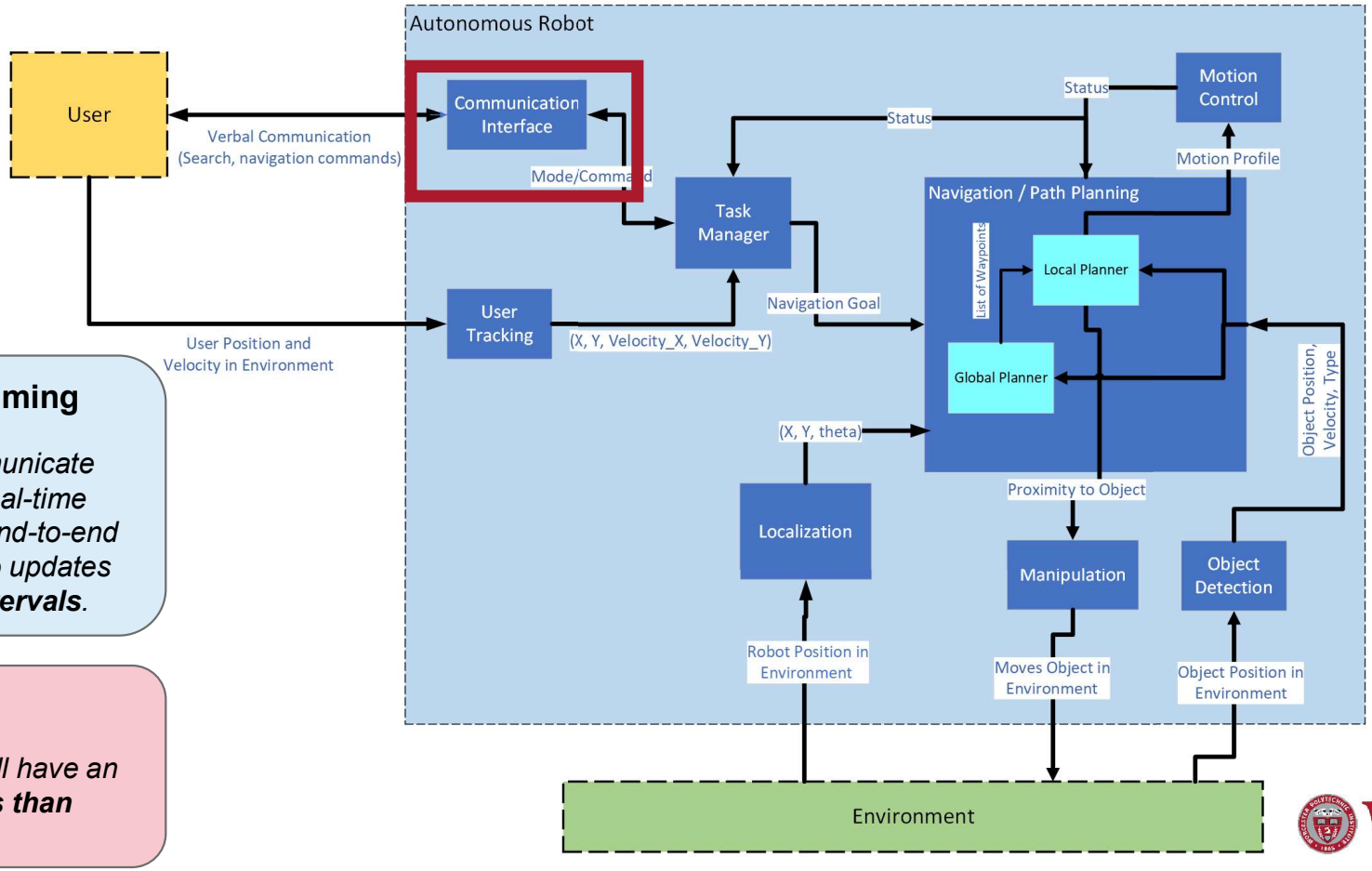
Communication System



Relevant Requirements

Communication Timing

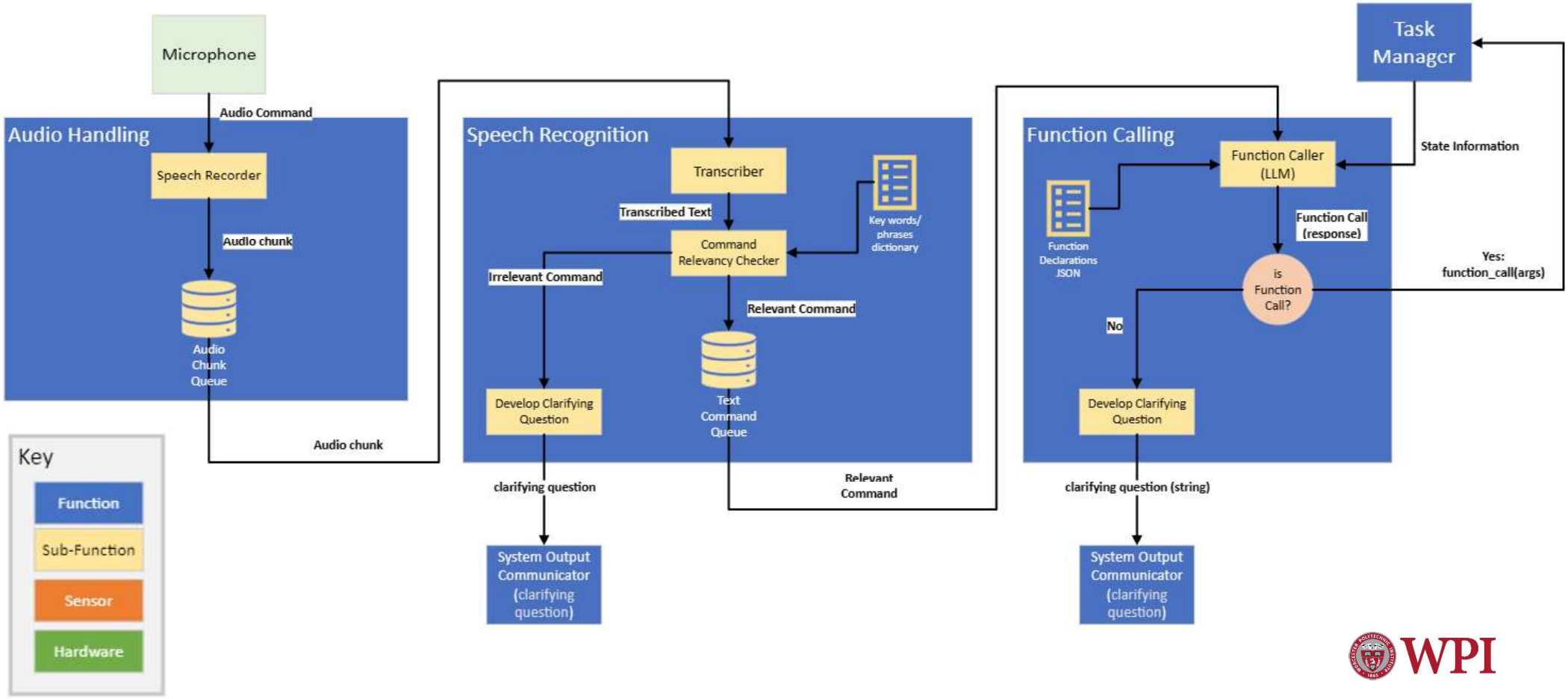
*The system shall communicate with the user in near real-time achieving a **30-second** end-to-end response time with audio updates at **most 10-second intervals**.*

Cost

*The proposed solution will have an estimated cost of **less than $50,000.***

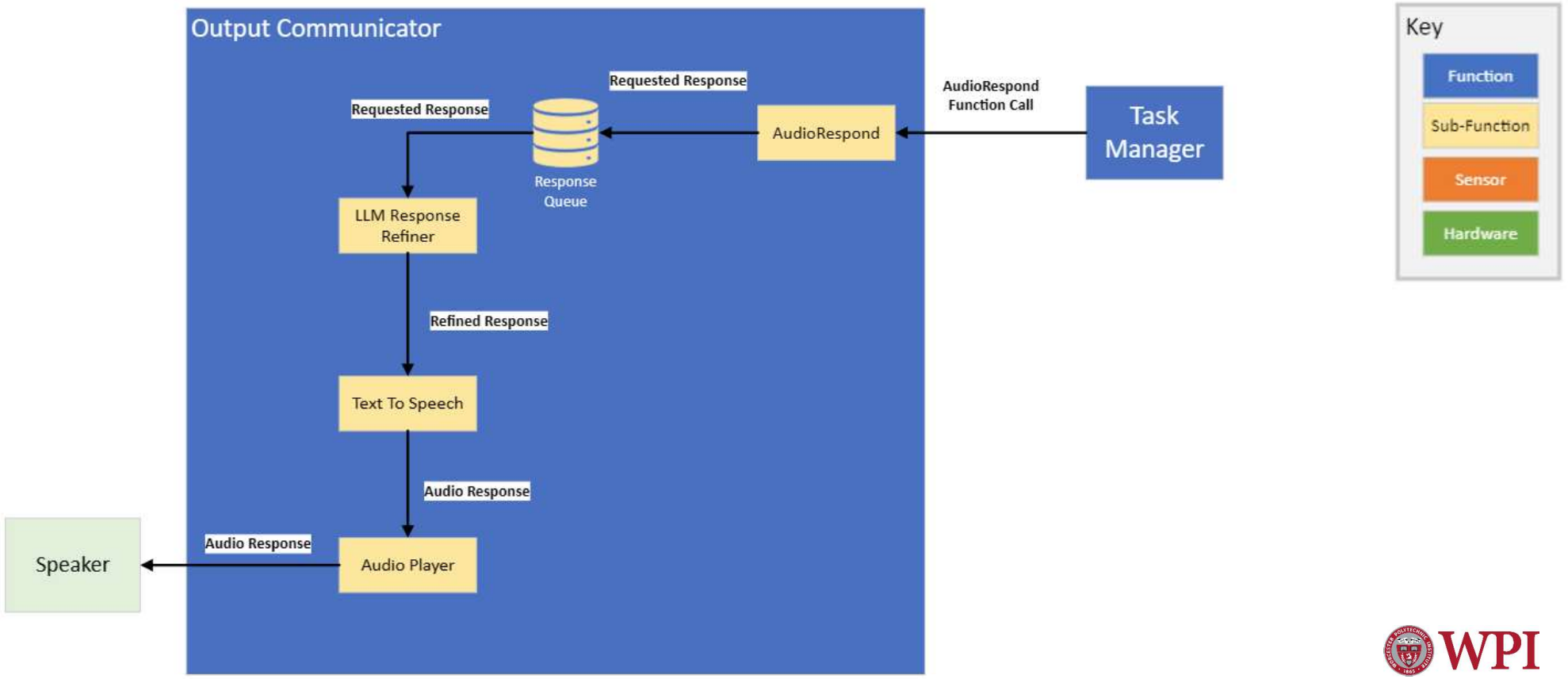
WPI

Communication System – Input



WPI

Communication System - Output

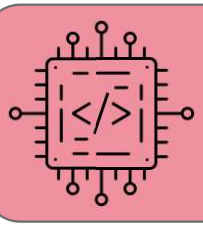


WPI

Communication System

Speech to Text

Response Refiner



Vector Embedding

Text to Speech

Function Calling

Location Refiner

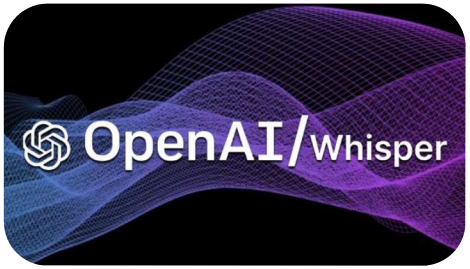
Input Function

Output Function

WPI

Communication System: Speech to Text

- Uses PyAudio mic stream with wakeword:
  - "Hey JIMBO": Joint Interactive Mobility Bot and Observer
- Speech Recognition module
  - Dynamically adjusts recorder for ambient noise
- Real-time transcriptions working using OpenAI whisper
  - Local instances were tested, but produced slow results with high word error rate (WER)
  - Uses English model to improve speed/accuracy



WPI

Communication System: Vector Embedding

- Reference documentation:
  - JSON architecture with system function descriptions and examples coded for semantic similarity
  - Sends to google embeddings model for vectorization
- Command relevancy subscriber
  - Picks up transcriptions from queue
  - Sends to embedding model for vectorization
  - Cosine similarity between embeddings
  - Thresholding to determine command relevancy
- If relevant, sends to function calling module
- Otherwise, respond to query and ask for clarification

WPI

Communication System: Function Calling

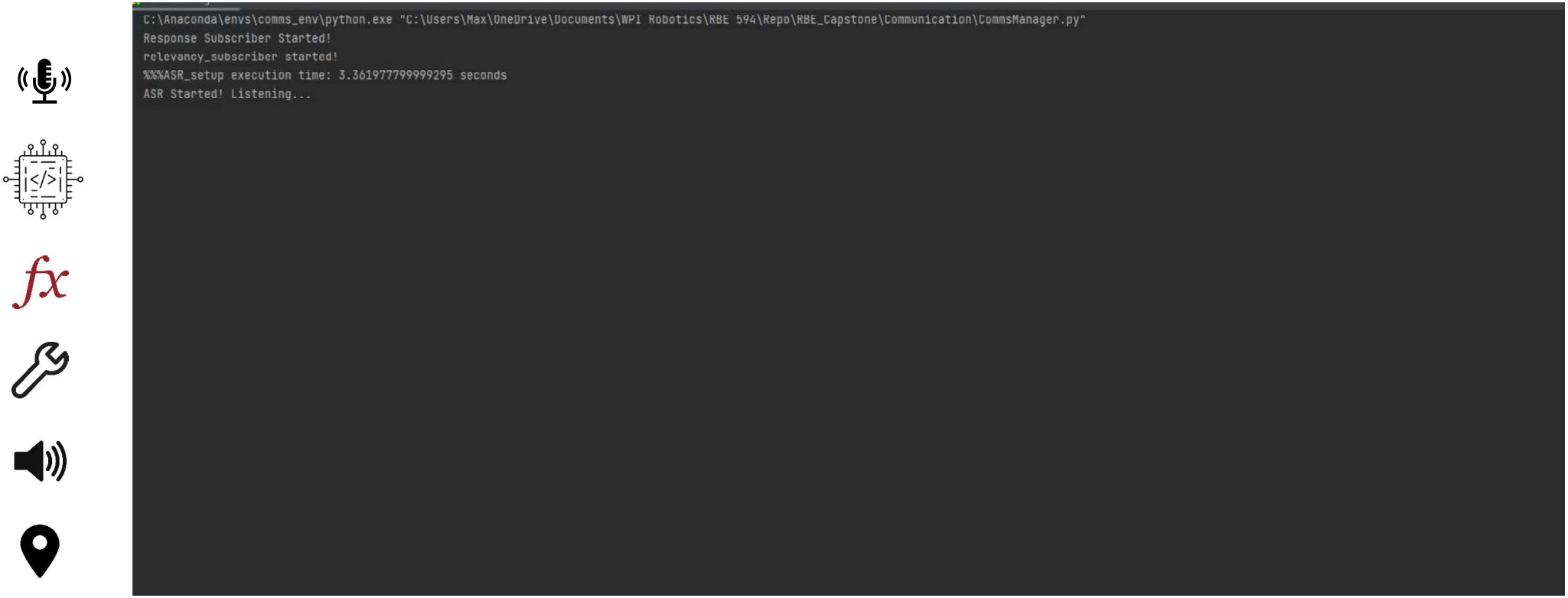
- Model: Google Gemini Pro
- Augmented with defined “tools”
  - Name
  - Definition
  - Parameters
- Response parsed for actual function call:
  - global_nav → location refinement
  - change_speed → motion controller
  - describe_env → image captioning (w/- Gemini Vision)
  - system_stop/system_go → motion controller

Function Set:

- **global_nav**
- **change_speed**
- **describe_env**
- **system_stop**
- **system_go**

WPI

Communication System: Function Calling Demo



WPI

Communication System: Response Refinement

- Implemented response refinement with context injection
  - Input: command + obstacle context
    - Ex: “Walk 10m, then turn right” + [[‘car’, ‘Left’], [‘pedestrian’, ‘Front Right’]]
  - Output: Refined response to user
    - Ex: “We're going to keep walking straight for a few more steps, and then we're going to be turning right. There is a car directly to our left and a pedestrian in front of us towards the right. I'll let you know before we need to turn."
- Model tuned on visual guide best practices
  - Support strategies for remote guides in assisting people with visual impairments for effective indoor navigation (Kamikubo et al., 2020)
  - Ex: Using clock syntax to describe orientation
  - Most important: providing context to instructions
- Originally used images as context
  - Created significant hallucinations

WPI

Communication System: Text to Speech

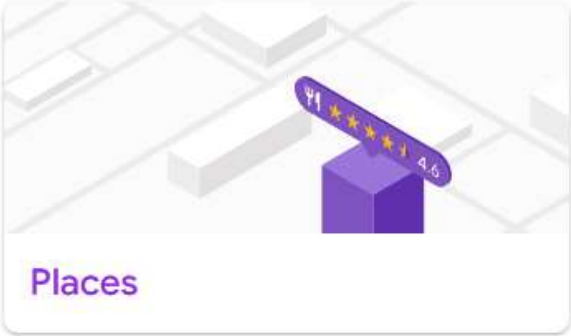
- Real-time text to speech (TTS) working using Eleven Labs
  - Built-in methods were tested, but more "human" voice was desired
- Using turbo v2 model for minimal latency
- Voice synthesis highly customizable

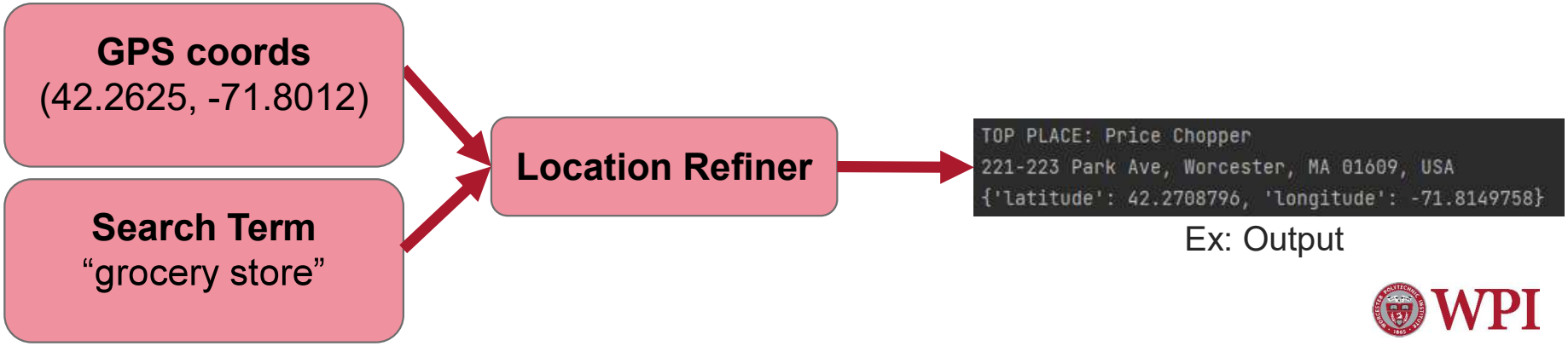
Eleven Labs

WPI

Communication System: Location Refiner

- Implemented using Google Maps Places API (new)
  - Takes in current GPS coords and search term
  - Assumes 1-mile search radius
  - Outputs top result with GPS location
  - Can include additional search params (i.e. accessibility)





Ex: Output

WPI

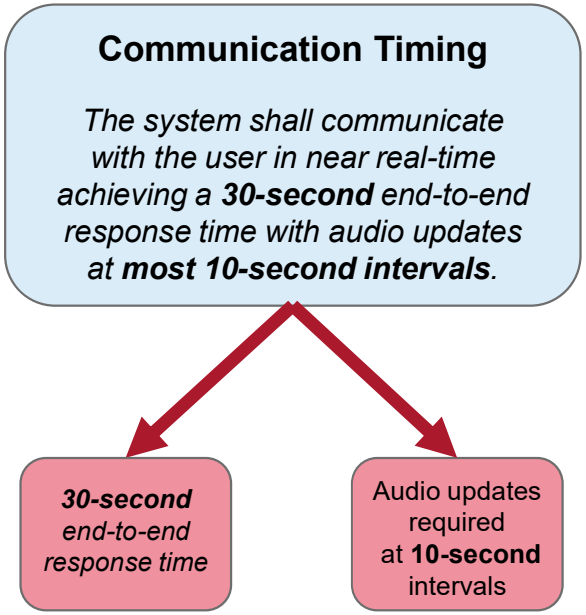
Timing Results

Module	Model Source	Avg Execution Time (s)
	Input Functions	
Speech to Text (Transcription)	OpenAI Whisper	0.838
Vector Embeddings	Google Embeddings	0.212
Function Calling – global_nav	Google Gemini	1.984
Function Calling – describe_env	Google Gemini	9.145
Function Calling - other	Google Gemini	1.568
INPUT TOTAL (avg function calling)		**5.282**
INPUT TOTAL (avg w/o describe_env)		**2.826**
	Output Functions	
Response Refinement	Google Gemini	1.837
Text to Speech	ElevenLabs	1.906
OUTPUT TOTAL		**3.743**
TOTAL (avg w/o describe_env)		**6.569**

WPI

Communication System: Conclusions

- 10-second audio updates:
  - Addressed through AudioResponse calls throughout system
  - Input Comms < 10s
    - describe_env not necessary, but “nice to have” (9.145s)
    - Avg input time w/o describe_env: **2.826s**
- 30-second end-to-end time:
  - Avg input + output time: **6.569s**
  - Response Refiner (RR) not necessary for all calls
    - Avg total time w/o RR: **4.732s**
- Timing results fall well within requirement threshold
  - Performance increases when only using core features

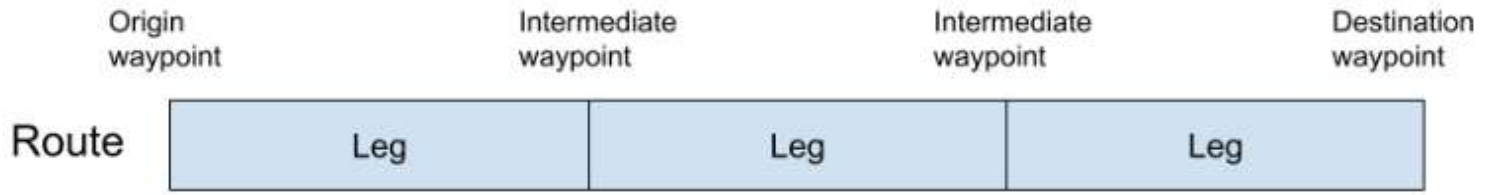


WPI

Global Planner

Global Planner Implementation

- Leverages Google Maps API
- Takes in starting location and destination
- Outputs walking route:
  - Discretized into legs separated by waypoints
    - Waypoints contain location information
    - Waypoints converted from GPS to ENU Cartesian coords
  - Waypoints then sent to local planner



WPI

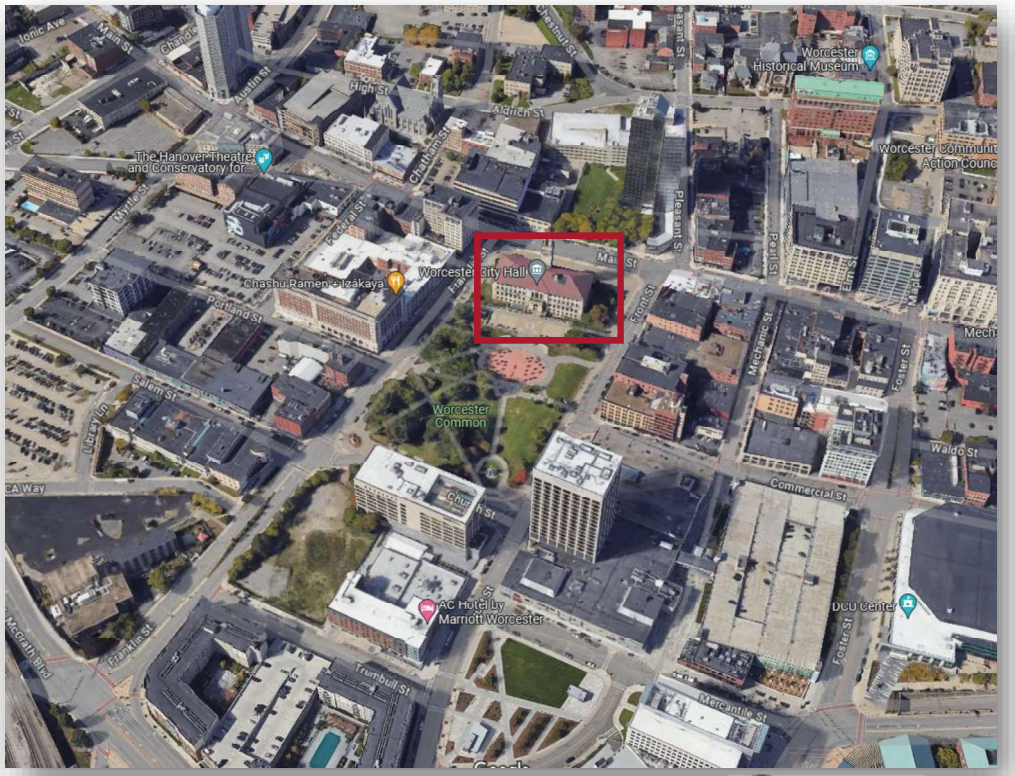
Simulation Results

Global Planner + Communication

Simulation World – Downtown Worcester

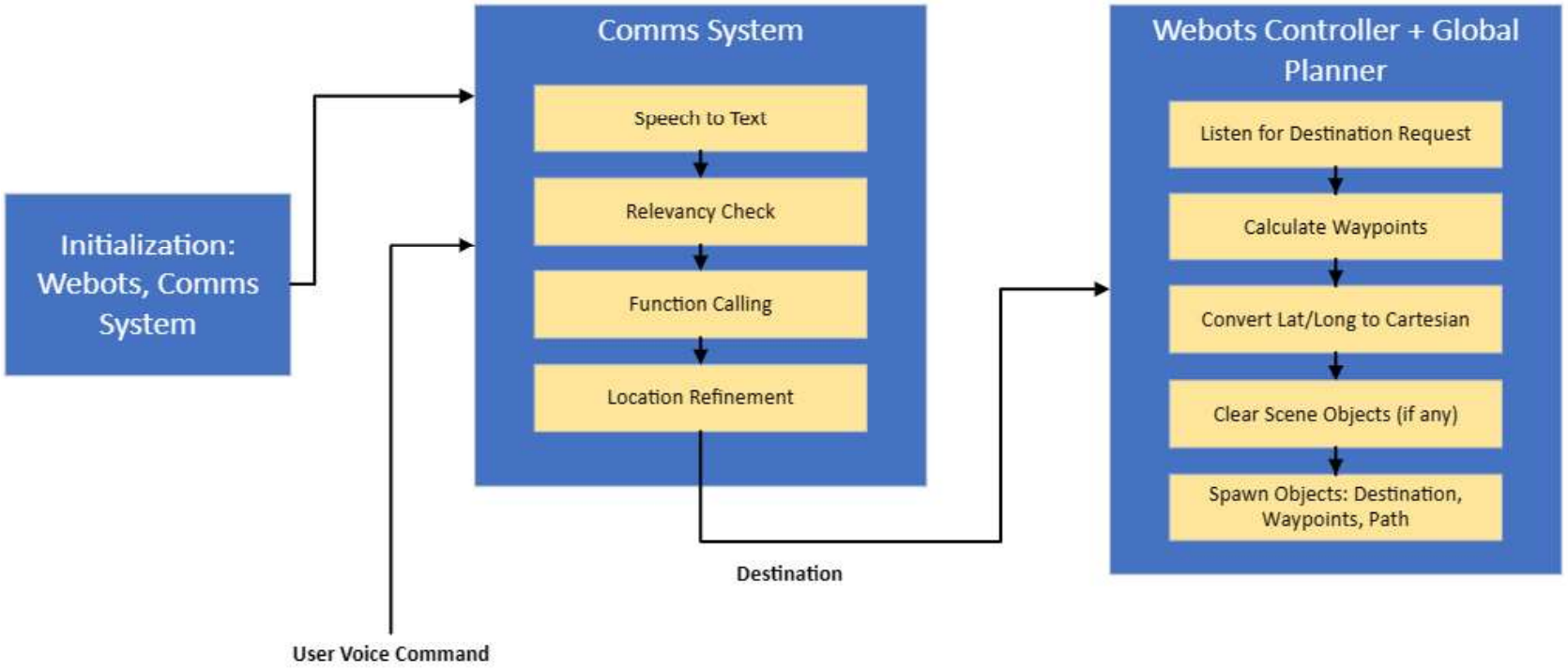
Webots Environment

Real World (Google Maps)



WPI

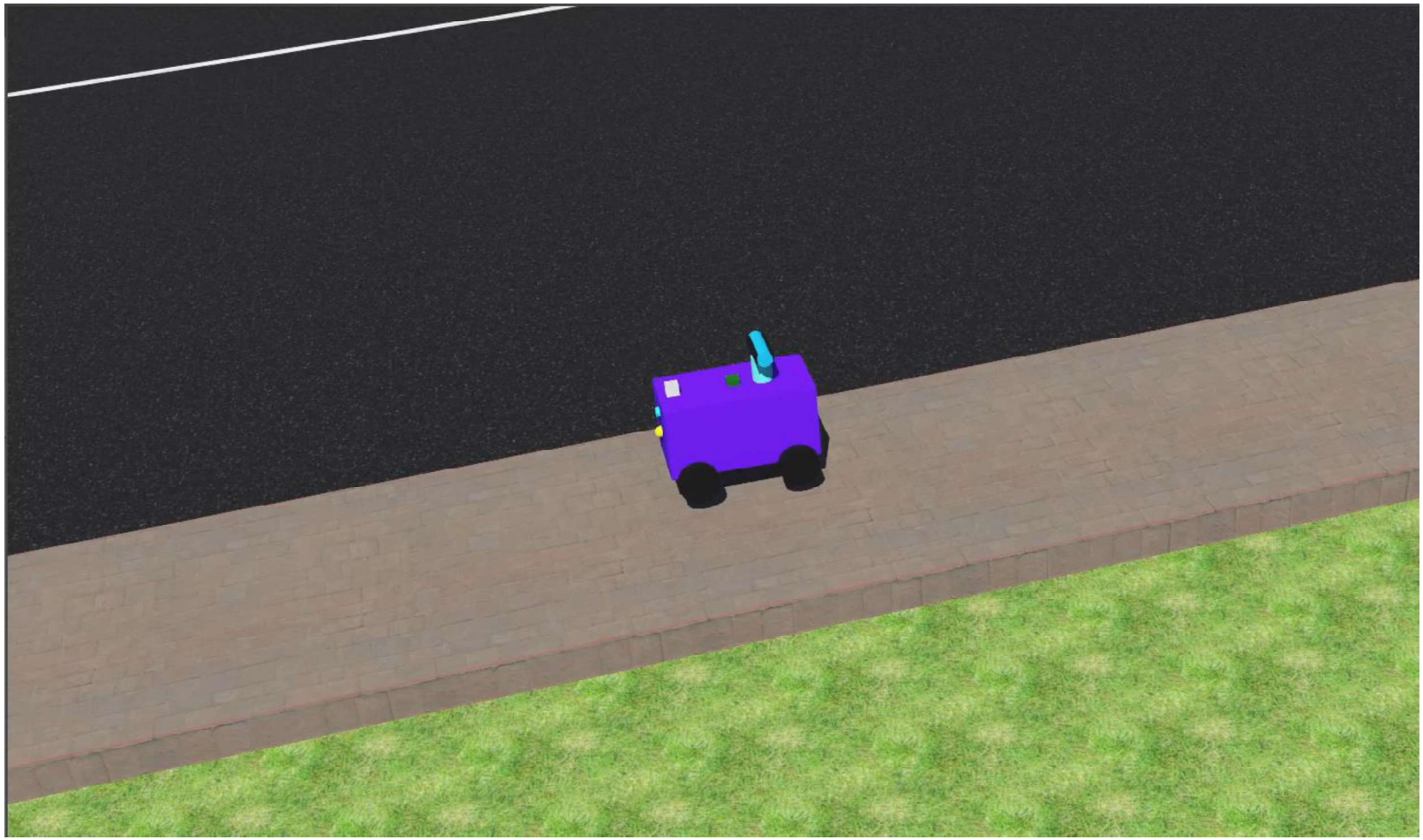
Simulation Flow

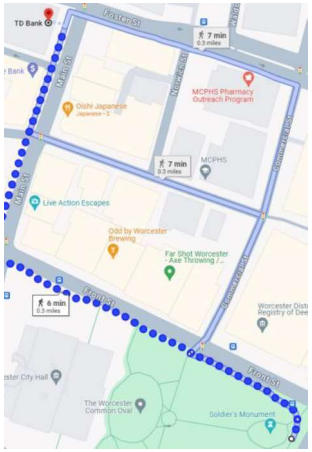


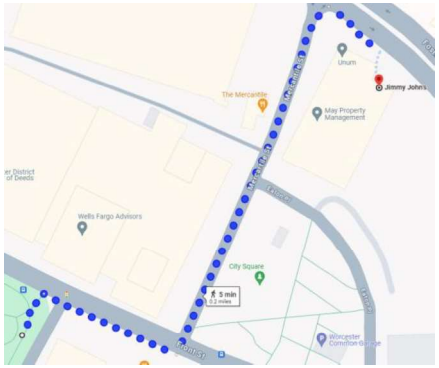
*Loops for each user input

WPI

Simulation Demo Video







WPI