AI Summary of Karim's Presentation

• Karim introduces the August 2025 DPRG meeting, himself as sleep-deprived, and Karim Virani's Pioneer 3 robot enhancements with an LLM.

• Karim introduces his robot, "Barney," based on a Pioneer 3 AT chassis, running ROS 2, and equipped with an **RTK GPS antenna** and a **Real Sense D455 depth camera** on a 4-degree-of-freedom arm.

• Discussion of the robot's **six-microphone array** with built-in hardware echo cancellation, which is crucial for preventing the robot from hearing its own voice.

• Karim outlines the user story: **fluid voice conversations** with the robot using frontier models, **understanding and carrying out objectives** (like a sentry robot patrolling property and interacting with people), and developing a **building understanding of the property** using photos and annotations into a graph RAG semantic database, all at a reasonable cost.

• Explanation of human conversational turn-taking (within 200 milliseconds) and challenges for robots, with a focus on **reducing latency** and the use of **real-time streaming models** like OpenAI's real-time API and Google's Gemini Live.

• Description of how real-time APIs send streams of camera, voice, and text, allowing LLMs to formulate responses before receiving the entire input and providing **streaming transcripts, voice responses, and text responses**.

• Karim discusses the need for an "agent" software to coordinate interactions and the **bifurcation of agents** (one for commands, one for conversation) due to differing conversation attributes and the inability to steer different output streams.

• Mention of Google's Gemini Live API and OpenAI playground for experimentation, noting **real-time models are about 10 times more expensive** than regular models.

• Clarification that the project focuses on high-level instructions (conversations, commands, behaviors, plans) and not fine motor control, introducing **Vision Language Models (VLMs)** and **Vision Language Action Models (VLAMs)** like Open VLA and Google Gemini's robotics initiative.

• Video demonstration of the Pi Zero robot folding laundry at 10x speed, highlighting the advanced capabilities of VLAMs.

• Comparison of OpenAI and Gemini APIs: OpenAI has better transcription but lacks real-time video, while **Gemini supports video** and is now preferred for Karim's project.

• The overall goal is to build a **sentry robot** that can detect intruders and engage in non-confrontational management.

• Discussion of how to explicitly manage the robot's listening (mute button) and the use of **Silero VAD (Voice Activity Detection) to stop streaming background sounds** to save costs and detect "go to sleep" commands, with a "double clap detector" to wake it up.

• Strategies for **cost optimization** with OpenAI using aggressive session cycling (killing sessions after utterances and reinjecting context) and noting that **Gemini is significantly cheaper** with generous free allowances.

• Demonstration of the robot processing a canned image, providing **JSON descriptions of objects with bounding boxes** in near real-time, using a simple prompt like "tell me what you see".

• Explanation of the **ROS (Robot Operating System) node graph** showing audio capture, VAD node, AI bridge, and agents (command and conversational) connecting to external servers.

• Live (though interrupted) **demo of Barney interacting**, identifying people and objects (like a traffic cone and fire extinguisher) in the room, and describing clothing.

• Discussion about using LLMs as **programming assistants**, emphasizing the need for high-level instructions and behaviors for the robot, and the extensive **logging and research notes** kept during development.

• Karim highlights the challenge of LLMs relying on outdated training data for new topics, requiring constant review and correction, and the benefit of **Claude Opus** for architectural decisions.

• The importance of **reinjecting conversational context** within a conversation (multiple sessions) and the limitations of the context window, leading to the need for **RAG (Retrieval Augmented Generation) and graph RAG** for persistent knowledge about the property and its changes.

• Karim confirms he is currently **focusing efforts on Gemini** and plans to share his GitHub repository to contribute to the community.