Demo Overview: Natural Language HMI

Purpose: Demonstrates a Human-Machine Interface (HMI) powered by a Large Language Model (LLM) executed on the Hailo-10H AI accelerator.

LLM Used: Qwen2 1.5B.

Hardware: Inference runs on the Hailo-10H AI accelerator, which is connected to a laptop using a Thunderbolt adapter. The GUI runs on the laptop.

Integration: Ollama APIs to interface with the LLM.

Key Features:

- Fully local inference – no cloud dependency.
- Simple and intuitive chat-based interface.

Demo Overview: Vision-Language Model

Purpose: Demonstrates a Vision-Language Model (VLM) capable of understanding and responding to combined visual and textual inputs.

LLM Used: Qwen2-VL.

Hardware: Inference runs on the Hailo-10H AI accelerator, which is connected to a laptop using a Thunderbolt adapter. The GUI runs on the laptop.

Key Features:

- Fully local inference – no cloud dependency.
- The VLM can recognize complex situations.

Demo Overview: Speech Recognition

Purpose: Demonstrates an Automatic Speech Recognition pipeline.

Model Used: Whisper (tiny/base).

Hardware: Inference runs on the Hailo-8 AI accelerator present on the System Electronics' Astrial SOM.

Key Features:

- Fast transcription.
- Computation offloaded to the accelerator.
- Low host CPU utilization.